ORSAY
n° d'ordre : 3341

UNIVERSITE DE PARIS - SUD

CENTRE D'ORSAY

THESE

présentée

Pour obtenir

Le TITRE de DOCTEUR 3ème cycle

Spécialité Statistiques

PAR

RAIS Noureddine

SUJET : APPLICATION DE LA THEORIE DES SONDAGES A UNE STRUCTURE RAMIFIEE
(à partir du recensement des oeufs d'hiver de P. Ulmi sur pommiers)

soutenue le 16 NOVEMBRE 1982 devant la Commission d'examen

MM. D. DACUNHA CASTELLE Président
J. BACHACOU
C. DUBY
G. FAUVEL
C. MILLIER

ORSAY

n° d'ordre : 3341

UNIVERSITE DE PARIS - SUD

CENTRE D'ORSAY

THESE

présentée

Pour obtenir

Le TITRE de DOCTEUR 3ème cycle

Spécialité Statistiques

PAR

RAIS Noureddine

SUJET : APPLICATION DE LA THEORIE DES SONDAGES A UNE STRUCTURE RAMIFIEE
(à partir du recensement des oeufs d'hiver de P. Ulmi sur pommiers)

soutenue le 16 NOVEMBRE 1982 devant la Commission d'examen

MM. D. DACUNHA CASTELLE Président
J. BACHACOU
C. DUBY
G. FAUVEL
C. MILLIER

UNIV.
LEIDEN
BIBL.

E R R A T A

p. 179 - 180 - 181 - 182 :

Changer ν par v

p. 184 : 4e, 5e ligne

Changer : " $\hat{B}_i$ parmi les estimateurs ξ-sans biais"

par : " $\hat{B}_i$ est optimal parmi les estimateurs ξ-sans biais de B_i"

avant dernière ligne : changer : "estimateurs ξ-sans biais"

par : "estimateurs linéaires ξ-sans biais"

p. 187 : 4e ligne : "conditionnellement aux B_i"

dernière ligne : dans l'expression de $S_{k\ell}$, changer "$i\, j \notin S$" par "$i\, j \notin S_\ell$"

p. 188 : Dans l'expression de $\hat{b}_\ell$ changer S par S_ℓ

p. 189 : Dans l'expression de $\hat{B}_k(\ell, A)$, changer " $\sum_{i \in A \cap S_k}$ par " $\sum_{i \in A \cap S_{k\ell}}$

Dans l'expression de T_ℓ, changer :

$$- \hat{b}_{0\ell} \quad \text{par} \quad \nu_{1\ell} \times \hat{b}_{0\ell}$$

$$- \frac{1}{n_\ell + {}_{1\ell}} \quad \text{par} \quad \frac{1}{n_\ell + \nu_{1\ell}}$$

R E S U M E

A partir du contrôle des pontes d'hiver dans un verger, est formulé le problème général de la stratégie d'échantillonnage dans une population ramifiée pour estimer la moyenne par arbre d'une variable mesurable sur tout segment, la mesure nécessitant de le prélever avec toutes ses ramifications.

Pour une structure régulière, ces propriétés de ramification et de prélévement sont traduites au niveau des probabilités d'inclusion du plan de sondage et au niveau du coût. La recherche des stratégies sans biais de la moyenne par âge des rameaux, est faite sous le modèle à population fixe puis sous 3 modèles de superpopulation retenus à partir des traitements des données. Des estimateurs sans biais de la moyenne par branche sont alors proposes et comparés par simulation dans le cas où la valeur de la variable sur un segment est liée par une relation de régression à celle sur ses ramifications.

M O T S C L E S

Echantillonnage - superpopulation - stratégie

Arbre - Ramification - Phytosanitaire

S U M M A R Y

Starting from the control of overwintering eggs in an orchard, the general problem of sampling strategy inside a ramified population is formulated with intent to estimate the variable's mean per tree. This variable can be measured in any segment, by cutting and drawing it with all its ramifications.

In the case of a regular structure, these characteristics are expressed in the inclusion probabilities of the sampling design, and in the total cost. The unbiased strategies of the mean per age are searched under the fixed population model and 3 superpopulation models dealing with the orchard data. Unbiased strategies of the mean per tree are then proposed and compared by simulation, in the case of a linear regression between the variable values upon one segment and upon its ramifications.

K E Y W O R D S

Sampling - Superpopulation - Strategy

Tree - Ramification - Pest Control

Je remercie Monsieur le Professeur Didier DACUNHA-CASTELLE de m'avoir fait l'honneur de présider le jury de cette thèse.

Que Monsieur Jean BACHACOU trouve ici l'expression de ma reconnaissance pour ses encouragements et sa participation tout au long de ce travail.

Je remercie Monsieur Claude MILLIER pour les conseils et les critiques utiles qu'il m'a prodigués et de l'honneur qu'il me fait.

Je remercie Madame Camille DUBY de l'intérêt qu'elle a bien voulu porter à mon travail en acceptant d'être membre du Jury.

Monsieur Guy FAUVEL a été l'instigateur de cette étude, m'a initié au problème biologique et m'a fourni les données nécessaires. Qu'il en soit vivement remercié ici.

Madame Josyane NEVEUX a assuré la dactylographie et le tirage de cette thèse avec son sourire et sa gentillesse habituelles. Je la remercie.

Tout le travail s'est passé dans l'ambiance chaleureuse et amicale de la Station de Biométrie du Centre National de Recherches Forestières à Nancy. Que tous ses membres trouvent ici l'expression de mon amitié.

Que toutes les personnes, ayant participé à la réalisation de ce travail soient assurées de ma reconnaissance.

INTRODUCTION

L'idée intuitive de ce travail est très simple. Pour l'expliciter nous allons prendre comme exemple les données traitées dans la partie A.

Une araignée s'attaque aux arbres fruitiers et cause des grands dommages à la récolte. Il existe des moyens de lutte contre cet acarien, qu'on applique suivant son "degré d'attaque". Pour prendre une décision (appliquer un traitement intensif, moyen ou pas du tout), on va chercher à avoir une idée, la plus exacte possible, sur le nombre d'araignées par arbre, ou plutôt sur la population en amont, formée par des oeufs déposés sur les branches.

Si ces oeufs sont assez gros pour être observés à l'oeil nu, on passe entre les arbres en examinant le plus possible d'organes. Pour éliminer "toute subjectivité", la personne (ou les personnes) effectuant ce sondage est sensée être guidée dans son choix par une méthode générale, appelée "plan de sondage", pré-établie selon nos connaissances a priori. Le risque de se tromper est alors d'autant plus faible que le nombre d'organes examinés est grand et que les connaissances a priori sont sûres.

Si par contre, ces oeufs sont difficilement discernables à l'oeil nu, il faudra alors couper les organes choisis et les examiner à l'aide d'un matériel approprié. Les dégâts occasionnés par ce prélévement peuvent alors être considérables. Or ces dégâts sont fonction des organes choisis (âge,...) ainsi que du plan de sondage (par exemple, prélever 2 branches de 5 ans sur un même arbre ou sur 2 arbres différents n'a pas les mêmes conséquences).

Comment ce problème était-il résolu jusqu'à présent ? On se fixe un âge, disons L, et on se ramène à un problème de sondage sur une population où chaque individu est une branche d'âge L. Pour un coût global fixé, on sait alors déterminer, à partir des connaissances préalables, un plan de sondage ainsi que le nombre de branches d'âge L à prélever. Cependant, la taille maximale de l'échantillon chute rapidement quand l'âge augmente. Alors, selon les

problèmes, soit on se limite à des âges tels que les dommages soient négligeables, soit on se suffit de peu de branches âgées. Dans les deux cas, le risque d'erreur peut être très grand pour des raisons différentes.

Nous proposons un plan de sondage qui permet de prélever un échantillon des branches qui, à la différence de la méthode courante, peuvent être de différents âges. Ceci peut être très important dans la pratique, surtout s'il s'agit d'un problème où il faut prélever des branches âgées, car un meilleur compromis peut être trouvé entre coût et taille de l'échantillon d'une part, et précision de l'estimation de l'autre. En effet, la structure des arbres fait que, dans nombre de problèmes, il existe une bonne relation (variable avec certains facteurs, par exemple le climat, le sol ...), entre la mesure sur un bois et celles sur ses ramifications (contagion, diffusion...). Ainsi, si le coût global nous permet de prélever 12 branches de 3 ans, nous proposons plutôt de prélever par exemple 7 branches de 3 ans, 2 branches en plus de 2 ans, et une soixantaine de rameaux de 1 an. L'utilisation des connaissances a priori au niveau de l'estimation nous permet alors de déterminer des estimateurs meilleurs que ceux employés jusqu'à maintenant.

Le présent travail se compose de 2 parties distinctes : la partie A traite du problème biologique du départ, le chapitre A-I présente le problème biologique et A-II présente brièvement les résultats des traitements de nos données.

La partie B traite du problème général d'échantillonnage sur une structure ramifiée.

Le chapitre B-I est la base de tout ce qui le suit, et présente le langage et la structuration utilisés pour décrire une population ramifiée. Dans B-II, le plan de sondage est défini par les probabilités d'inclusion d'ordre 1 et 2 ; les propriétés de la population (ramification) ainsi que celles de nos prélévements (si un rameau est prélevé, alors tout ce qu'il porte l'est aussi) sont traduites à ce niveau.

Pour B-III, IV et V, le problème de l'estimation de la moyenne sur les branches d'un certain âge L sur tout le verger est traité dans des cas particuliers correspondants à certains résultats de la partie A. Finalement,

nous nous intéressons surtout au cas où une bonne relation de régression lie des mesures sur un rameau à celles sur l'ensemble de ses ramifications. Un estimateur est alors proposé dans V-6, et son comportement est étudié par simulation dans certains cas particuliers.

Il est sûr que nous n'avons fait que poser le problème. Beaucoup reste encore à faire, notamment par exemple à utiliser les techniques du sondage séquentiel en tenant compte de l'âge, à améliorer nos estimations en définissant des modèles décrivant nos connaissances a priori sur les valeurs des paramètres (modèle bayésien), ect...

Etant donné la diversité des sujets traités, nous avons préféré citer la bibliographie à la fin de chaque chapitre. Cependant, les deux livres de base dans toute la partie B sont :

- *W.G. COCHRAN : Sampling Techniques*
- *C.M. CASSEL, C.E. SARNDALL, J.H. WRETTMAN : Foundations of inference in survey sampling*

Nous avons suivi la démarche de ce dernier dans l'estimation (B-III, IV et V).

S O M M A I R E

Pages

INTRODUCTION

SOMMAIRE

PARTIE A : Recensement des oeufs d'hiver de Panonychus-Ulmi (koch) sur pommiers

A-I - PRESENTATION 1

1) Biologie de l'araignée rouge (P.Ulmi) 3

2) Méthodes utilisées 7

Bibliographie 12

A-II - TRAITEMENT DES DONNEES 13

1) Programmes utilisés 14

2) Présentation - Notations 17

3) Premières observations 23

4) Analyse factorielle des correspondances 29

5) Régressions 33

6) Analyse de variance et de covariance 51

7) Conclusions 75

Bibliographie 78

PARTIE B : Application de la théorie des sondages à une structure ramifiée

B-I - POPULATION 83

Bibliographie 94

B-II - PLAN DE SONDAGE 95

1) Notations et définitions générales - Introduction 97

2) Contrainte d'inclusion 102

3) Probabilités d'inclusion d'ordre 1 105

4) Probabilités d'inclusion d'ordre 2 109

5) Liens entre plan de sondage et probabilités d'inclusion 123

Bibliographie 128

B-III - INTRODUCTION A L'ESTIMATION 129

1) Observations générales 130

2) Modèle à population fixe 134

3) Modèle de superpopulation 139

4) Fonction coût 142

Bibliographie 146

B-IV - MODELE A POPULATION FIXE 147

1) Choix d'une stratégie 148

2) Propriétés de la stratégie (P, T_{gd}) 152

Bibliographie 159

B-V - MODELE DE SUPERPOPULATION 161

1) Modèles considérés à partir des traitements des données 162

2) Modèle 1, stratégies P-sans biais 165

3) Stratégie ξ-sans biais - Modèle M1 169

4) Stratégie ξ-sans biais - Modèle M2 174

5) Modèle M3 178

6) Utilisation des mesures effectuées sur U_h, $h < k$, *comme variables auxiliaires dans l'estimation de* $\bar{Y}_k$ 183

Bibliographie 191

B - VI - SIMULATION 192

1) Description des programmes 193

2) Résultats 201

3) Conclusion 213

Bibliographie 214

CONCLUSION 215

ANNEXE DE A-II

ANNEXE DE B-IV

P A R T I E A

RECENSEMENT DES OEUFS D'HIVER DE PANONYCHUS - ULMI (KOCH) SUR POMMIERS

A - I - PRESENTATION

Dans 1), des éléments de la dynamique de l'araignée rouge Panonichus-Ulmi (Koch) sont présentés brièvement ainsi que l'intérêt du recensement des oeufs d'hiver et quelques éléménts sur leur répartition.

Dans 2), nous indiquons les méthodes utilisées pour l'estimation des effectifs de la population hivernale de P.-Ulmi, ainsi que les dernières voies de recherche.

G. FAUVEL

STARKING

MANDON

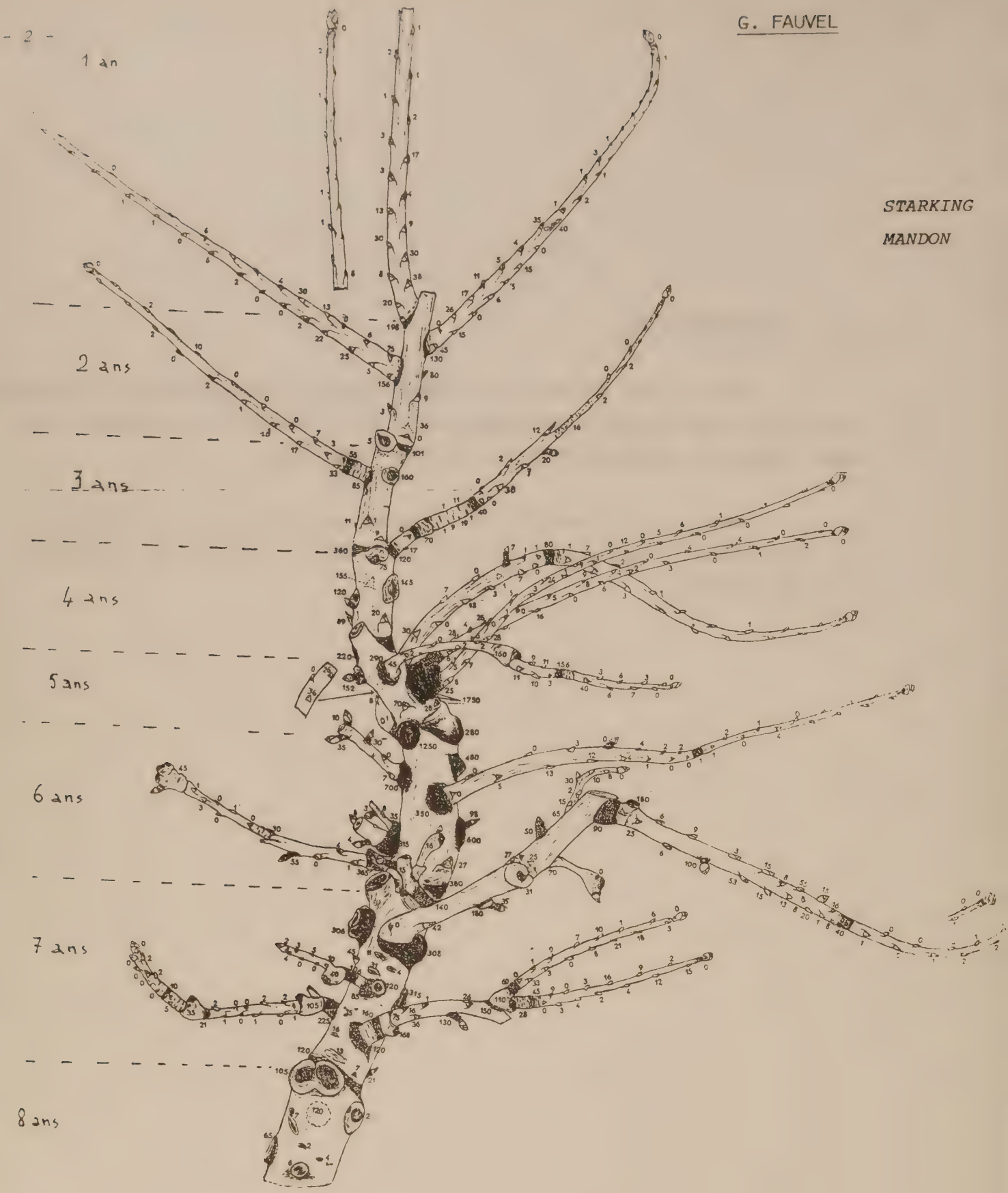


FIGURE 1

A-I-1) BIOLOGIE DE L'ARAIGNEE ROUGE (PANONYCHUS ULMI KOCH)

L'araignée rouge est un acarien de dimension très réduite qui se nourrit des feuilles des arbres fruitiers (pommiers, poiriers, pruniers, vigne,...).

L'adulte femelle mesure 0,4 mm de long, il est de forme globuleuse rougeâtre. Le mâle est plus petit, allongé, de couleur beige pâle.

L'araignée rouge hiverne à l'état d'oeufs dits "oeufs d'hiver", de forme presque sphérique, rouge brique, le diamètre pouvant atteindre 0,13 mm.

Leur diapause n'est rompu que s'ils ont subi des températures inférieures à 5°C. Le cas échéant au printemps, ils donnent naissance à une première génération constituée uniquement de femelles. Leurs pontes ne donneront naissance qu'à des mâles.

La nouvelle génération s'accouple avec la première pour donner une population mixte. Ainsi se chevauchent de 5 à 7 générations (le nombre est fonction de la température).

Son développement est complexe, l'acarien passant par 7 formes : l'oeuf, la prélarve, la larve, deux stases mobiles, une stase immobile et l'adulte.

Les conditions climatiques (températures, hygrométrie de l'air) influencent très fortement l'évolution de l'acarien et peuvent l'anéantir complétement.

On estime de 12 à 18 jours la durée de vie des femelles dont la ponte (24 à 48 oeufs) est fonction de l'état physiologique du feuillage.

Trois facteurs influent sur la ponte d'hiver : la photopériode, la température, et la nature de l'alimentation.

Cette ponte s'effectue en général de fin août à fin septembre à raison d'un oeuf par jour et par femelle.

Ech : 1/6,8

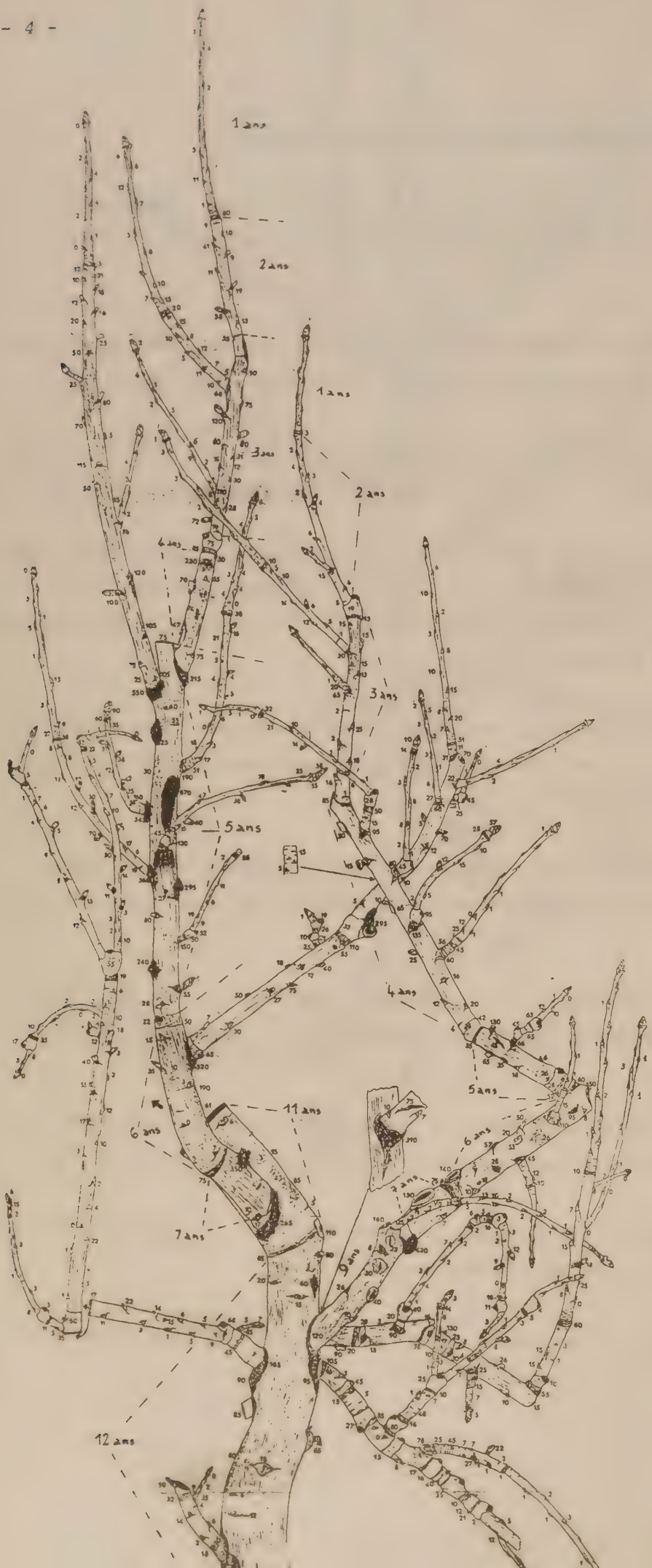


FIGURE 2

La pullulation de l'acarien rouge est un phénomène récent, survenu il y a une vingtaine d'années, à la suite de l'utilisation des insecticides de synthèse, et aggravé ensuite avec l'emploi des fongicides organiques.

L'alimentation par le feuillage ainsi traité provoque chez l'acarien une augmentation de longévité et de fécondité ainsi que sa résistance aux divers acaricides.

- DEGATS

L'araignée rouge s'attaque au tissu palissadique des feuilles, ce qui donne une couleur plombée à la face supérieure.

Les femelles issues des oeufs d'hiver se concentrent sur les jeunes feuilles. Les dégâts se manifestent principalement en juillet, période où la pullulation est maximale.

Ces attaques entrainent une régression de la photosynthèse, et même le dessèchement et la chute des feuilles. Elles se traduisent par une baisse sensible de la qualité et de la quantité de la récolte.

Son potentiel élevé de reproduction fait de Panonychus Ulmi un des plus importants ravageurs des vergers industriels.

- INTERET DU RECENSEMENT DES OEUFS D'HIVER

A partir de l'estimation de l'effectif de la population hivernante, un traitement peut être décidé, d'où une action sur la population en amont pouvant conduire à la limitation des populations successives.

Il est intéressant de remarquer qu'une quantité importante des oeufs d'hiver est détruite naturellement (jusqu'à 40 % et au-delà), et qu'il est préférable d'effectuer les prélévements le plus tardivement possible.

Cette période correspond en général à celle où la taille des arbres est effectuée. D'autre part, ceux-ci sont encore sans feuilles, ce qui minimise le coût et facilite le comptage.

ECH : 1/8

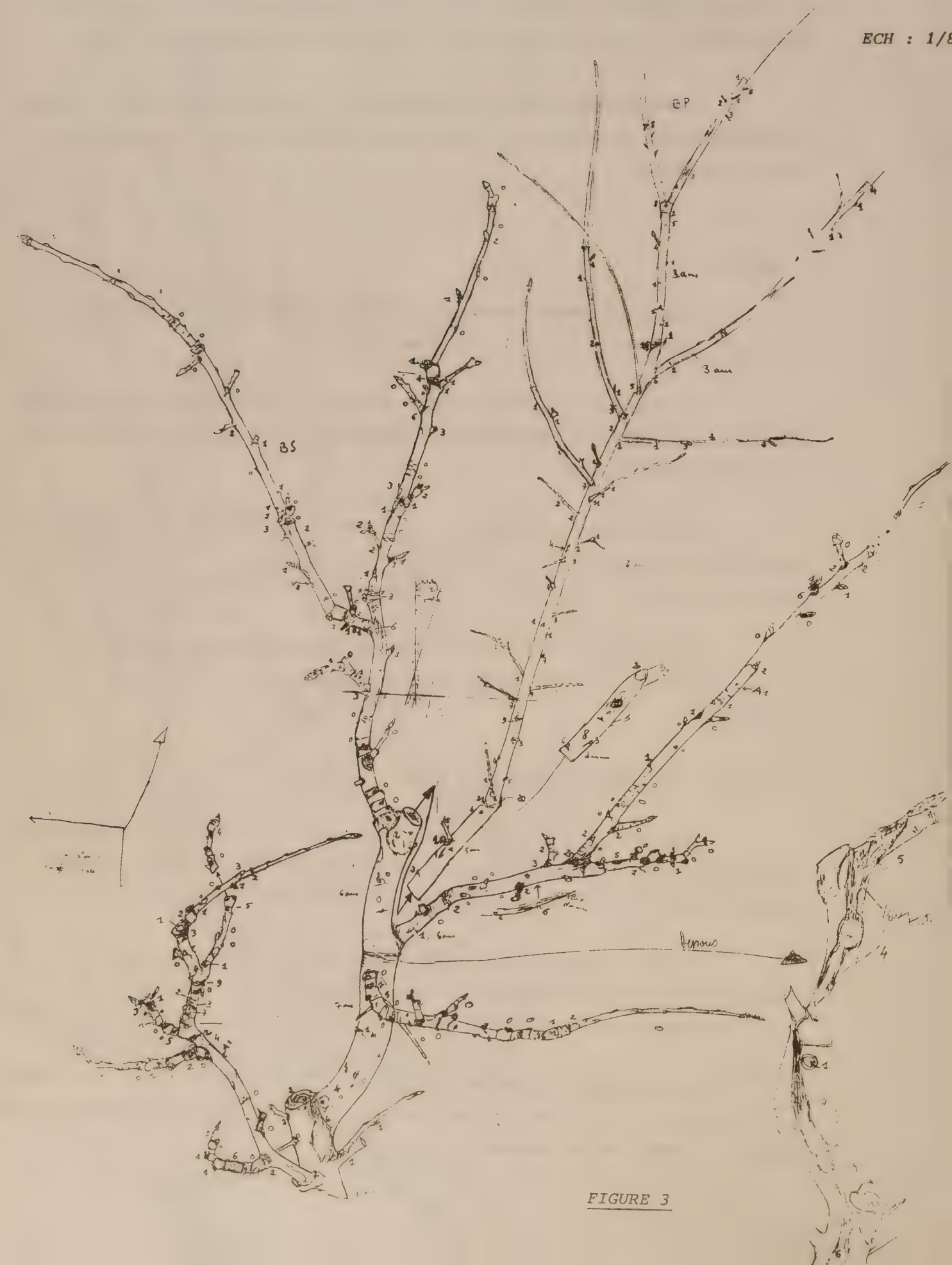


FIGURE 3

Pour effectuer les comptages, on expose les rameaux à un puissant jet d'eau qui détache les oeufs ; ceux-ci sont alors recueillis dans un filtre facilitant l'opération.

. Facteurs influençant la répartition des oeufs [2]

- Topographie du verger (présence ou absence de pente)
- Orientation (préférence sud = soleil)
- La forme de l'arbre et sa position dans le verger
- Variété fruitière
- L'exposition aux vents

. Les oeufs sont déposés de préférence [3]

- sur la face inférieure
- à des endroits rugueux, sur des obstacles ou caches potentielles

A-I-2) METHODES UTILISEES

Pour l'estimation des formes mobiles, D.P. PIELOU [5] a établi que le pourcentage $\frac{n_0}{N}$ de feuilles non infestées par l'araignée rouge est lié au nombre moyen M d'acariens par feuille par la relation $E(\frac{n_0}{N}) = \exp(-M)$.

Des méthodes simples et rapides sont utilisées pour l'estimation du degré d'infestation par les formes mobiles.

Pour l'estimation du nombre d'oeufs d'hiver, plusieurs méthodes existent, se basant sur l'observation de bois de 1 an à 3 ans. La méthode de référence a été redéfinie par VOGEL et BACHMAN en 1969 (OILB) comme suit :

- effectuer 10 prélévements par arbre, sur 5 arbres dans le verger. Un prélévement est défini comme un morceau de bois de 20 cm de longueur de 2 à 4 ans, de préférence parmi les rameaux rugueux et richement pourvus de dards et de bourgeons à fleurs. Le dénombrement des pontes est effectué sur tout l'organe, dards, lambourdes et empattements de l'année compris.

Cette méthode demande trop de travail pour une faible précision. Néanmoins, toutes les autres méthodes s'y réfèrent pour établir leur admissi-

G. FAUVEL

CH : 1/6,8

FIGURE 4

bilité et la majorité n'en sont qu'une simplification.

FAUVEL [3], a comparé les méthodes dérivées à la méthode originelle sur 14 vergers.

- OILB simplifiées :

. Le prélévement consiste en 20 cm de bois sans ramification : lambourdes, dards, bourgeons, brindilles. L'empattement terminal n'est pas compris.

- BAILLOD - FIAUX (1975)

. Le dénombrement ne se fait plus que sur 2 organes consécutifs et opposés sur le bois de 2 ans. (dards, bourgeons bien développés, ou petits rameaux adjacents).

- 10 arbres au moins par verger
- 100 prélévements par verger (soit 10 par arbre en moyenne)

Ces 2 méthodes présentent l'avantage d'être plus pratiques, mais leur représentativité est bien moins sûre que la méthode originelle.

- Empattements de l'année

Cette zone reçoit un fort pourcentage des pontes sur les rameaux de 1 et 2 ans.

- cinq arbres par verger
- 10 prélévements par arbre
- un prélévement correspond à l'empattement du bois de 1 an.

- Touzeau

- 10 arbres par verger
- 10 prélévements par arbre
- un prélévement correspond au premier couple d'yeux développé à la base de la pousse de l'année.

- Méthode des 10 yeux

Le comptage se fait sur 1/4 à 1/3 des bourgeons du rameau.

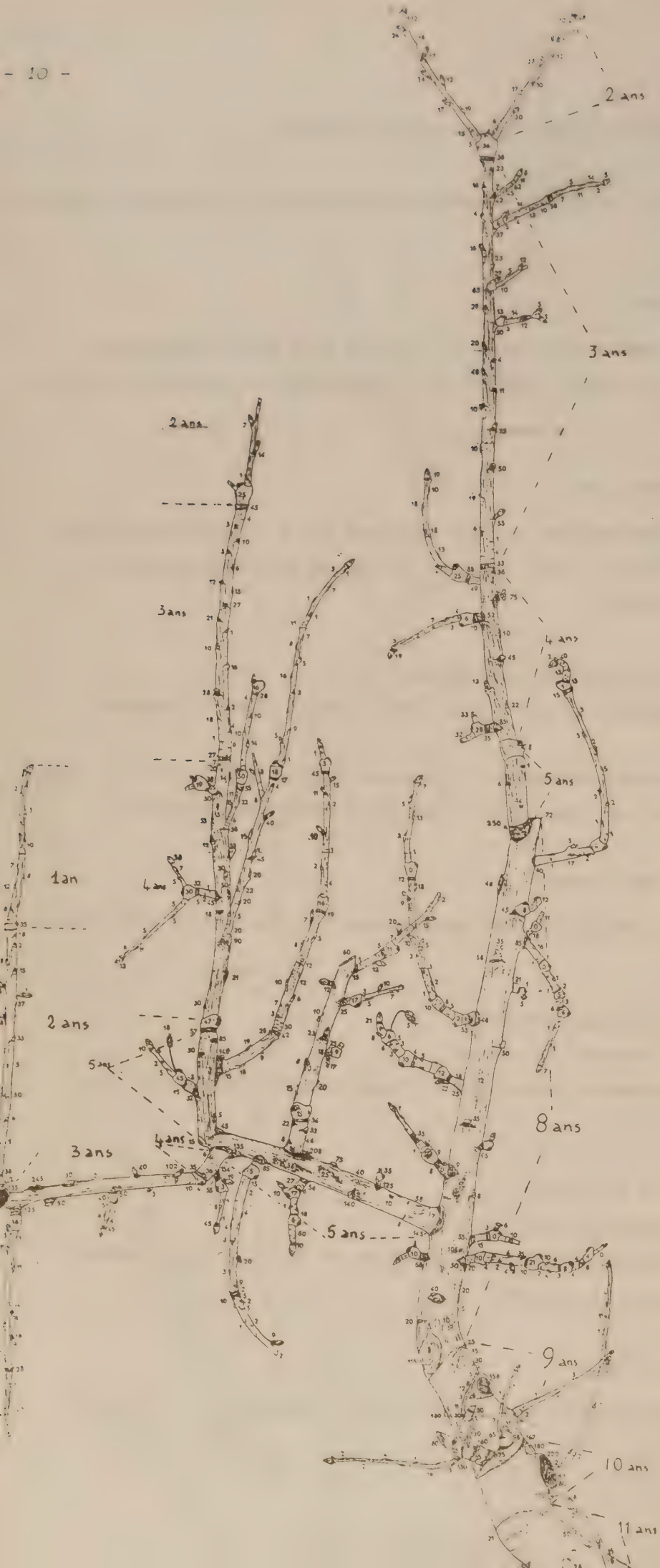

2 ans
3 ans
2 ans
3 ans
4 ans
5 ans
1 an
2 ans
5 ans
3 ans
4 ans
8 ans
5 ans
9 ans
10 ans
11 ans

FIGURE 5

Toutes ces méthodes sont très corrélées avec la méthode OILB de référence (le coefficient de corrélation varie entre 0.78 et 0.99). Les méthodes de BAILLOD - FIAUX et OILB simplifiée ont les plus faibles variations [3].

L'avantage de toutes ces méthodes c'est qu'elles sont pratiques et simples. De plus, en ne considérant que les bois de 1 an à 3 ans, les dommages causés au niveau de l'arbre et du verger sont négligeables. Mais toutes ont aussi une faible précision quant à prévoir l'infestation estivale. En effet la relation entre l'importance des dépots des oeufs sur les jeunes rameaux et celles des populations estivales est très mal connue, il en résulte une incertitude dans la fixation des seuils de traitement.

Alors que pour les formes mobiles, il est compréhensible de se limiter à un examen des feuilles puisque les araignées s'en nourrissent, il n'y a aucune raison a priori de ne considérer que les bois de 1 à 3 ans pour l'estimation des effectifs d'oeufs d'hiver.

Dans un travail récent, GASSER et KLAY [4] étudient la répartition des oeufs d'hiver sur 4 demi palmettes. Ils arrivent à la conclusion suivante "il est ainsi impossible de compter les oeufs sur bois de 1 an (ou de deux ans) et d'en déduire la quantité totale" ([4], p.41).

En désignant par cache "une surface délimitée, par exemple, un oeil, un départ de pousse, un empattement, une fissure, une lenticelle plus profonde, une cicatrice restant après la cueillette des fruits", GASSER et KLAY montrent que sur les 4 demi-arbres, "si [la densité absolue sur un arbre] augmente, le taux d'occupation des caches fait de même et ceci quelque soit le type de bois considéré".

Signalons finalement une idée intéressante de SONDEY [6] qui met en relation le nombre d'oeufs d'hiver de P.ULMI sur 20 cm de rameau de 2 à 3 ans et la fréquence d'occupation des 2 premiers bourgeons des rameaux de 1 an.

Nous formulons cependant un doute sur la portée de la majorité des travaux et la possibilité de généralisation des résultats du point de vue statistique, lesquels se suffisent d'un faible nombre d'observations.

BIBLIOGRAPHIE

[1] A C T A SPECIAL 5 "VIGNE" - Avril 1975

[2] P.J. BUSSCHERE (1970)

Distribution des formes hivernantes des ravageurs du pommier dans la région de Gembloux.
(Novembre et décembre 1967)

Bull. Rech. Agron. Gembloux - N.S. TOME V - n° 1-2

[3] G. FAUVEL, H. AUDEMARD, F. RODOLPHE, A. RAMBIER (1978)

Le recensement des oeufs d'hiver de l'acarien rouge, Panonychus Ulmi (Koch) sur le pommier.
I - Comparaison des méthodes actuelles et remarques sur l'influence de quelques facteurs dans leur variabilité

Ann. Zool. Ecol. Anim., 1978, 10(3), p.461-482.

[4] R. GASSER et A. KLAY (1980)

Etude de la répartition des formes hivernantes d'acariens phytophages en vergers de pommiers et éléments de la dynamique de leurs populations en automne.

Travail effectué à l'Institut d'emtomologie de l'E.P.F.Z. sous la direction du professeur V. DELUCCHI.
Semestre 79-80.

[5] D.P. PIELOU (1960)

Contagious distribution in the european red mite, Panonychus Ulmi (Koch), and a method of grading population densities from a count of mite-free leaves

Can.J.Zool. 1960, (38), P. 645-653

[6] J. SONDEY (1976)

Comparaison des méthodes de prognose hivernale utilisée sur vigne et sur pommier pour déterminer les seuils de nuisibilité concernant les oeufs d'hiver de Panonychus-Ulmi.

Observ.Trav. 1975, Serv. P.V. Fr., 6, 67-73

[7] TOUZEAU (1973)

La répartition des oeufs d'hiver de Panonychus Ulmi sur arbres fruitiers - Propositions concernant la prognose hivernale.

ACTA, note d'inform., 11, p.5-9

A - II - TRAITEMENT DES DONNEES :

Les données traitées ont été fournies par M. FAUVEL (Laboratoire d'Ecologie Animale et de Zoologie Agricole - I.N.R.A. MONTPELLIER).

Dans 1), nous faisons quelques remarques sur les programmes utilisés.

Dans 2), une présentation des données.

Dans 3) à 6), les plus importants résultats sont exposés. D'autres traitements n'ayant rien donné n'ont pas été rapporté.

A-II-1) PROGRAMMES UTILISES

Dans la majorité des cas, les programmes utilisés sont ceux de la programmathèque AMANCE [1].

Les notations utilisées sont très répandues, signalons néanmoins que le résultat du test d'une hypothèse H_0 sera indiqué par '×' (resp. '×××'). Si H_0 est rejetée avec un risque de première espèce de 5% (resp. de 1%), l'absence d'étoiles indiquant que H_0 n'a pas été rejetée.

Nous indiquons sommairement la façon d'établir des résultats non fournis par les programmes d'AMANCE :

- le programme de la régression progressive REPRG ([ ], p.209) fournit $\hat{b}$ et S^2 estimateur des moindres carrés de b et σ^2 dans le modèle :

$$\underset{nx1}{Y} = \underset{nxk}{X}\underset{kx1}{b} + \underset{nx1}{\varepsilon}$$

avec ε de loi $N_n(O,\sigma^2 I_n)$

Soit $C : R^k \to R^{k-h}$ une application linéaire. Le domaine de rejet au niveau α de l'hypothèse $\{Cb = Cbo\}$ contre $\{Cb \neq Cbo\}$ est défini par :

$$D\alpha = \{\frac{(\hat{b}-b_0)'C'(C(X'X)^{-1}C')^{-1}C(\hat{b}-b_0)}{(k-h)\ S^2} > f_{\alpha,k-h,n-h}\}$$

Le calcul de cette statistique se réduit au calcul d'une somme de carrés sur tous les individus.

- Pour homogénéiser la présentation des résultats, nous avons calculé à la suite du programme AVFAC ([1],p.301 : analyse de variance multidimensionnelle d'un dispositif factoriel orthogonal), le coefficient de corrélation multiple R ainsi que le test global d'ajustement F

$$1 - R^2 = \frac{\text{Somme des carrés résiduels du modèle}}{\text{Somme des carrés totale}}$$

$$F = \frac{R^2/k}{(1-R^2)(n-k-1)}$$ *avec n le nombre d'observations et k la somme des degrés de libertés de toutes les composantes du modèle*

- De même, après le programme COVAR ([1],p.459, analyse de covariance), dans le cas où une seule variable est expliquée. La statistique de Wilks vérifie alors :

$$\frac{1}{U(1,n-k,k-h)} - 1 = \frac{k-h}{n-k} F(k-h, n-k)$$

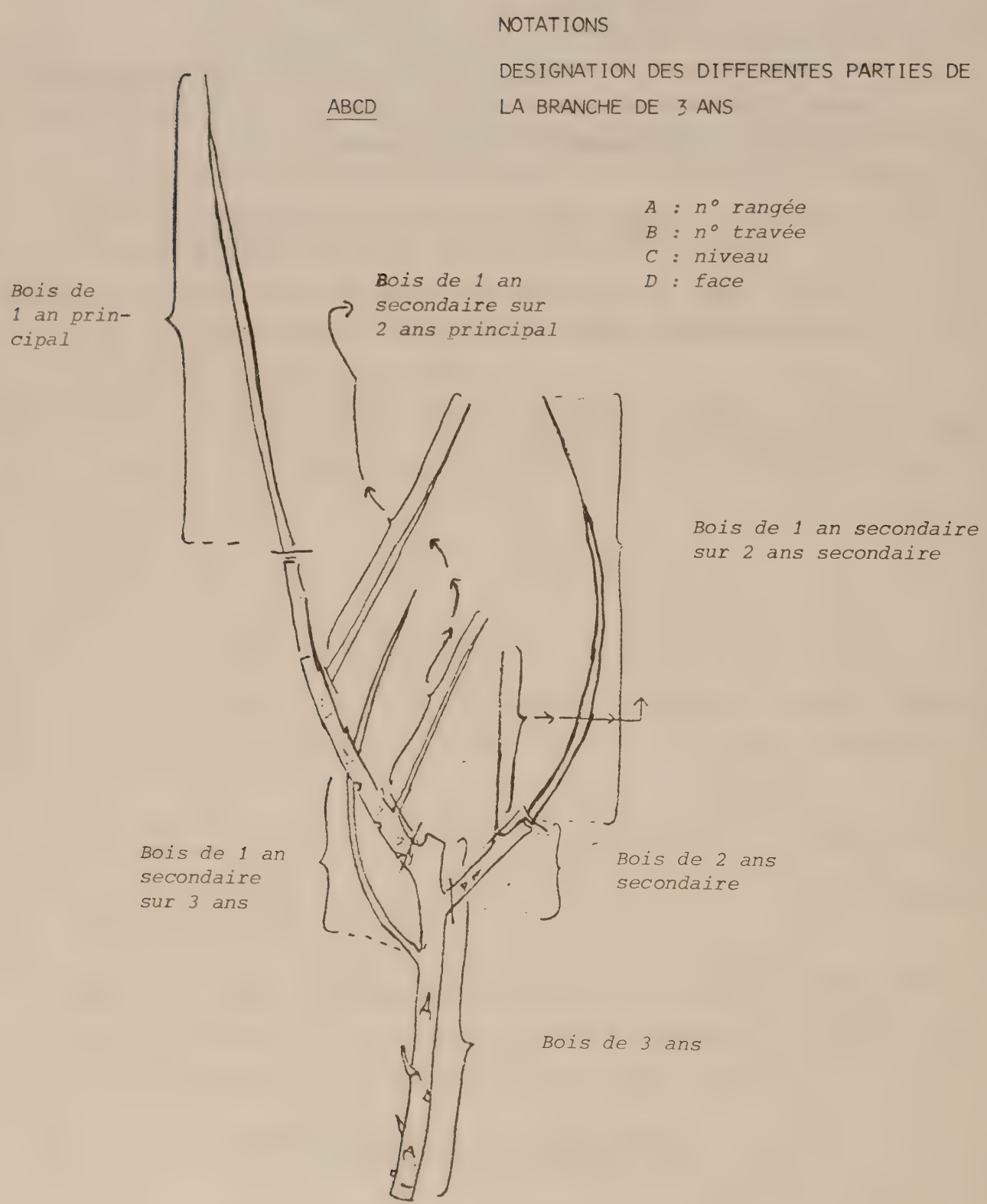
si p est le nombre de variables explicatives et k les degrés de libertés du modèle d'analyse de variance associé, alors le test de signification d'une variable X_i a pour région de rejet de niveau α :

$$D\alpha = \{(\frac{SC2}{SC1} - 1) \cdot (n - k - p-1) > F_{\alpha,1,(n-k-p-1)}$$

avec SC1 (resp. SC2) la somme des carrés résiduels dans le modèle avec les p covariables (resp. dans le modèle avec p covariables sauf X_i)

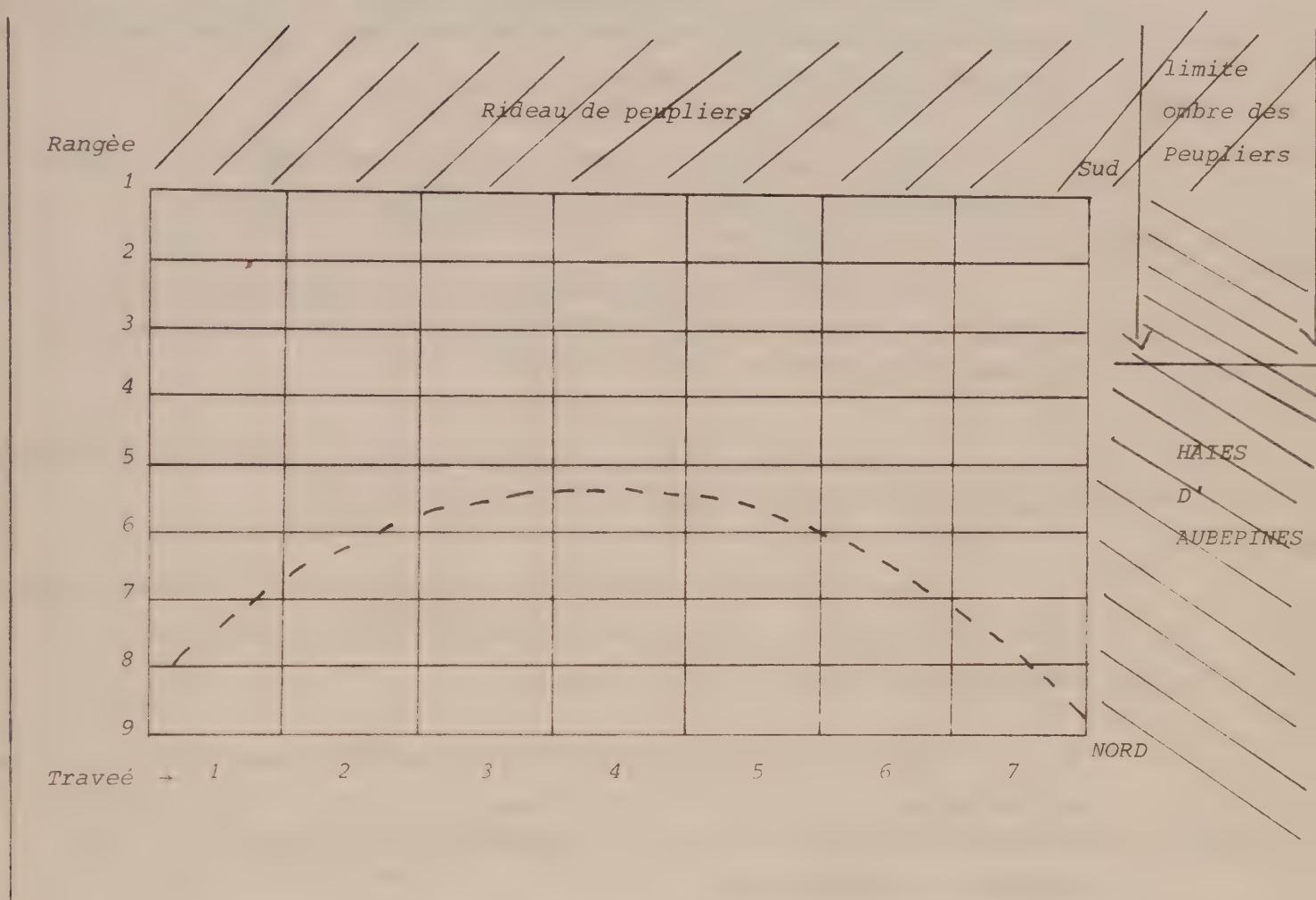
Un autre programme donné en annexe a été utilisé pour le calcul des statistiques de BARTLETT.

FIGURE 6

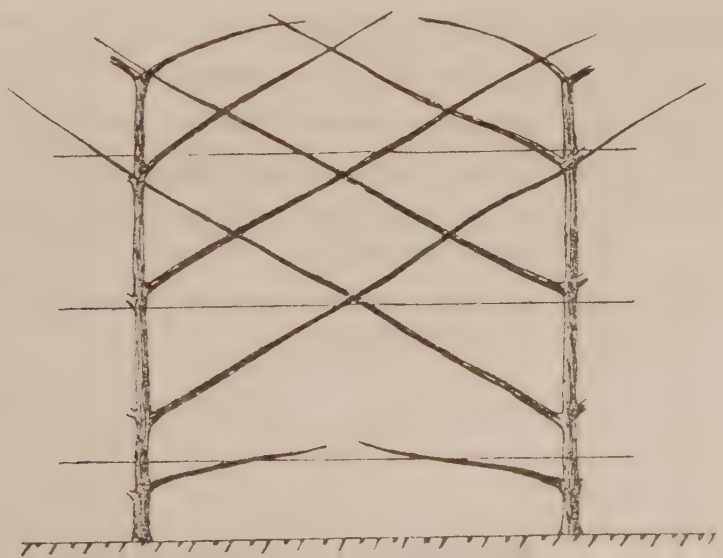

NOTATIONS
DESIGNATION DES DIFFERENTES PARTIES DE LA BRANCHE DE 3 ANS
ABCD
A : n° rangée
B : n° travée
C : niveau
D : face
Bois de 1 an principal
Bois de 1 an secondaire sur 2 ans principal
Bois de 1 an secondaire sur 2 ans secondaire
Bois de 1 an secondaire sur 3 ans
Bois de 2 ans secondaire
Bois de 3 ans

A-II-2) PRESENTATION - NOTATIONS

Les données ont été prélevées dans l'hiver de 77-78, sur un verger de pommiers de la variété Golden, dans la région de Montpellier (la Barthelasse). La position du verger est comme suit :



Le verger est constitué de neuf rangées, chacune comprenant sept haies formées suivant la méthode dite Palmettes obliques :



Chaque haie est constituée en moyenne de sept arbres, et nous supposerons que la structure (architecture) globale des haies est peu variable dans le verger.

Remarquons que l'ombre des peupliers couvre les trois premières rangées. Les rangées sont bien séparées (4 à 5 métres), alors que la structu des haies lie bien les arbres d'une rangée entre eux et ne nous permettra pa de raisonner au niveau de l'arbre.

Pour chaque haie, nous avons trois niveaux codés 1, 2 et 3 pour bas, moyen et haut et deux faces Nord (codée 1) et Sud (codée 2).

Dans chaque haie, pour chaque niveau et chaque face, une branche de 3 ans est prélevée au hasard, soit 6 branches par haie et 378 branches sur tout le verger.

Le plan d'échantillonnage est systématique, équilibré à 4 facteurs rangée, travée, niveau et face.

Chaque branche de trois ans est décomposée en six parties définies selon l'âge (1, 2 ou 3 ans), l'importance (principal, secondaire) et le rameau porteur (fig.1).

Nous appellerons population sur un rameau le nombre d'oeufs d'hiver de Panonychus - Ulmi déposés sur ce rameau. Le terme "caractéristiques" sera réservé aux variables décrivant l'architecture de la branche telles que longueur, inclinaison, etc...

La population sur une partie de la branche sera notée :

POP3 : pour le bois de 3 ans

P2P : pour le bois de 2 ans principal

P2S : pour la totalité des bois de 2 ans secondaires

P1P : pour le bois de 1 an principal

P1SP : pour la totalité des bois de 1 an secondaires portés par le bois de 2 ans principal

P1SS : pour la totalité des bois de 1 an portés par l'ensemble des bois de 2 ans secondaires

P1S3 : pour la totalité des bois de 1 an portés directement par le bois de 3 ans

P2T : pour la totalité des bois de 2 ans (soit P2T = P2P + P2S)

P12PT : pour la totalité des bois de 1 an portés par le bois de 2 ans principal (soit P12PT = P1P + P1SP)

P2PT : pour la totalité des bois de 1 an porté par le bois de 2 ans principal, et le bois de 2 ans principal (P2PT = P2P + P12PT)

P12T : pour la totalité des bois de 1 an portés par tous les bois de de 2ans (soit P12T = P12PT + P1S3)

P1T : pour la totalité des bois de 1 an (P1T = P12T + P1S3)

PTOT : désignera le nombre d'oeufs déposés sur toute la branche :

$$PTOT = POP3 + P2T + P1T$$

un M à la place du T indique la population moyenne : P2M est donc la population sur l'ensemble des rameaux de 2 ans portés par une branche.

Les longueurs (en cm) seront notées similairement LONG3, L2P, L2S, L1P, L1SP, L1SS, L1S3, L2T, L12PT, L2PT, L12T, L1T et LTOT.

Notons aussi :

PBRS : la population totale sur toutes les bourses du bois de 3 ans

PBRSM : PBRS/nombre de bourses sur le bois de 3 ans

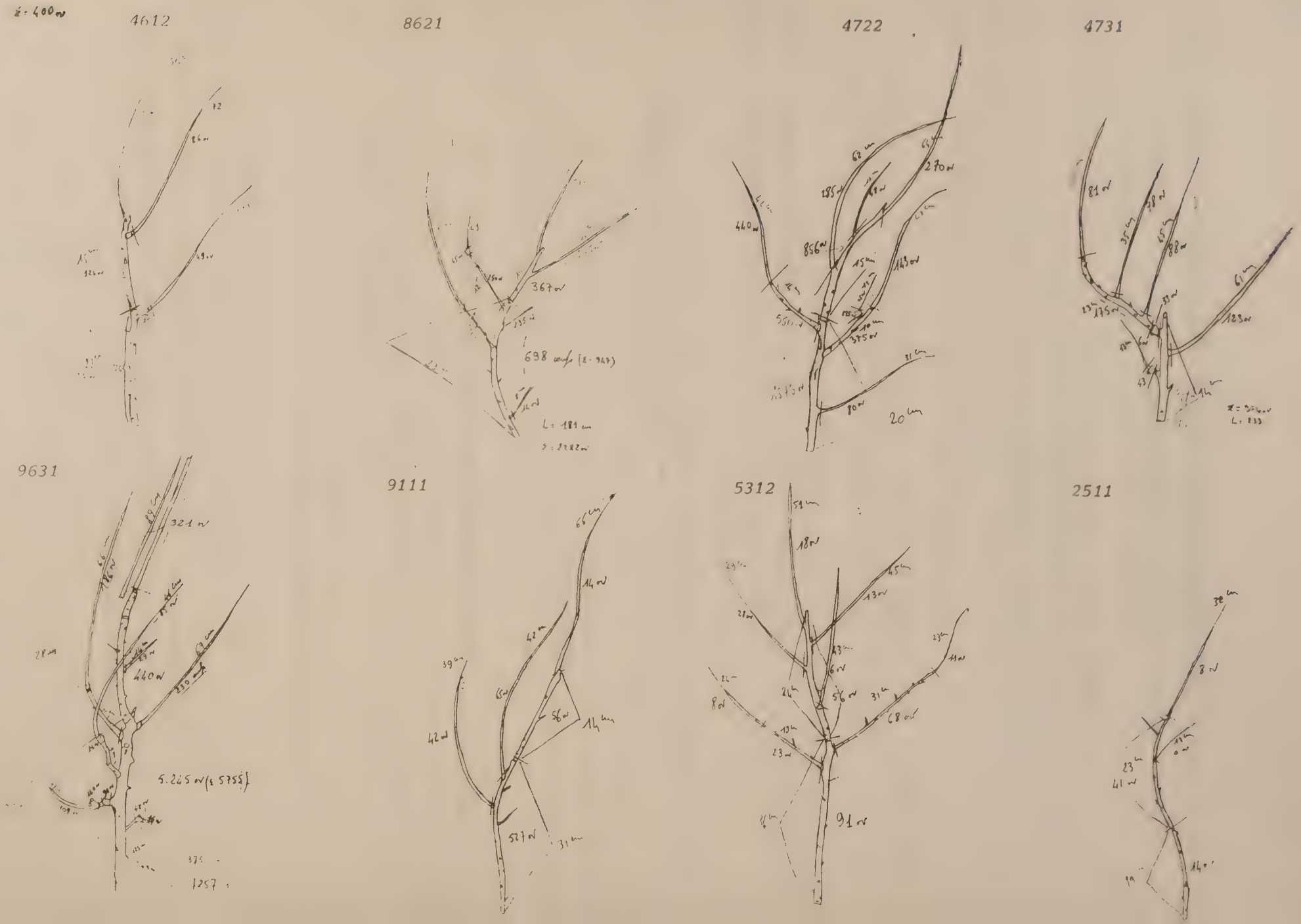
Un L sera rajouté à toutes les notations précédentes pour indiquer que la transformation $x \to log(x+1)$ a été appliquée à la variable :

$$LP2T = log(P2T + 1)$$

$$LL2P = log(L2P + 1)$$

D'autres caractéristiques de la branche ont été utilisées, mais se sont avérées sans intérêt dans la suite telles que les inclinaisons (verticales), l'orientation cardinale (angle que fait la projection du rameau sur le plan horizontal avec le nord). Aussi, nous ne nous encombrons pas des notations correspondantes.

FIGURE 7 : Architecture de quelques branches



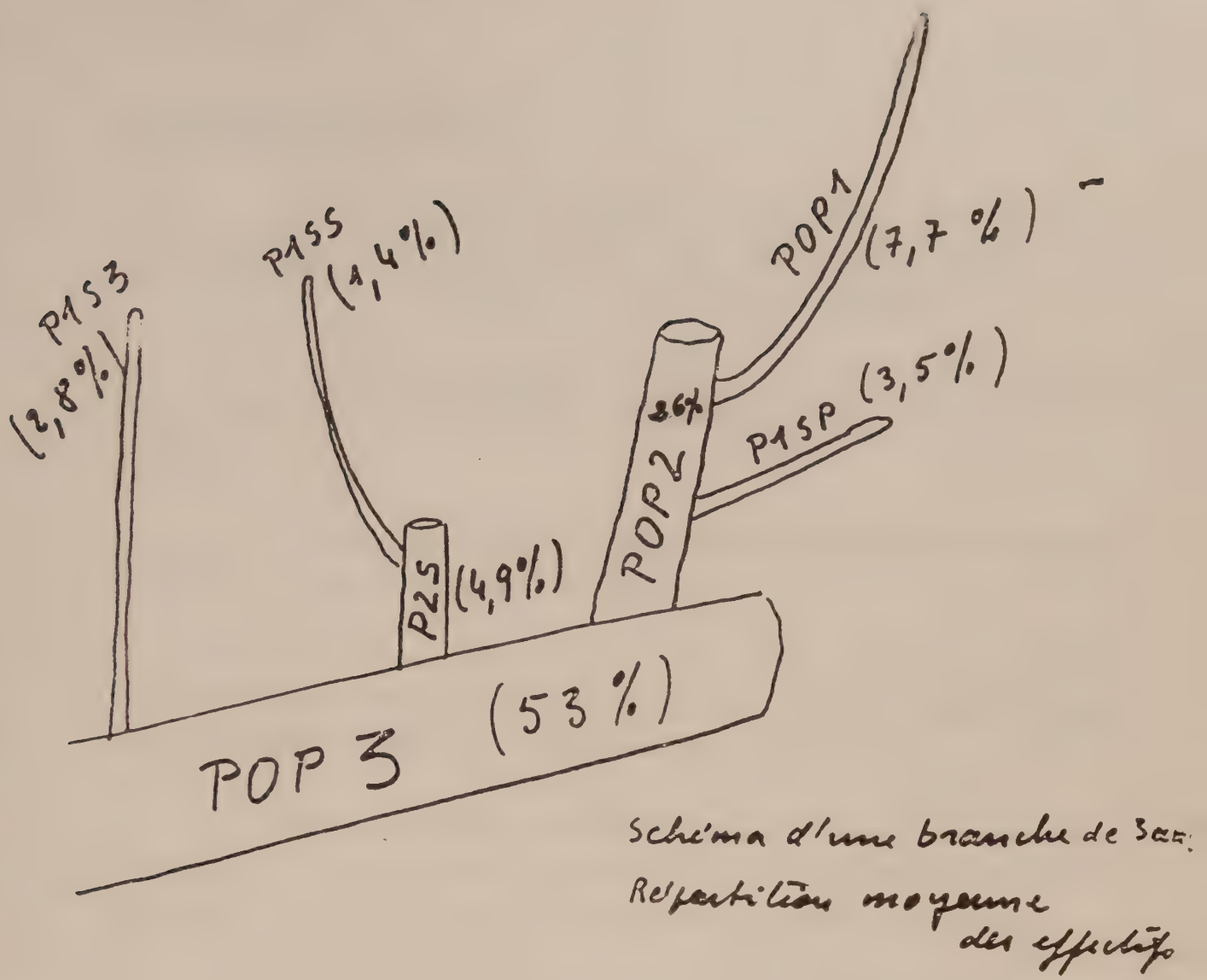


Figure 8 : *Pourcentage moyen d'oeufs sur les différentes parties d'une branche*

A-II-3) PREMIERES OBSERVATIONS :

- L'architecture des branches est très variable (fig.2). Par exemple, le nombre de rameaux de 2 ans sur le bois de 3 ans varie de 1 à 4, et celui des 1 an portés par le bois de 2 ans principal de 0 à 5.

En conséquence, nombre des données sur les variables P2S, P1SS, P1S3, P1SP et PBRS sont manquantes (les longueurs correspondantes sont nulles).

- Le pourcentage moyen d'infestation des différentes parties de la branche est représentée dans la fig.8.

Mais ceci est très variable d'une branche à l'autre. La figure 9 représente les populations sur quelques branches particulières, ainsi que l'allure moyenne sur toute la population (ligne discontinue)

- les statistiques élémentaires des variables sont données dans le tableau T1. Nous constatons qu'en général, les coefficients de variation des populations sont proches de 1 sauf pour P2S, P1SS, P1S3 et PBRS.

Comme dit plus haut, ces quatres populations sont particulières : nous avons seulement 199 branches, parmi les 378 considérées, telles que L 1S3 non nul, 124 pour L 1SS et 87 pour PBRS.

Dans le cas général, si X est une variable de moyenne m et d'écart type f(m), et s'il existe g telle que g'(x) = 1/f(x), alors la moyenne et la variance de g(x) sont indépendantes.

Nous avons donc considéré la transformation X → Log(x + 1). La valeur 1 a été rajoutée pour éviter le problème des zéros, mais n'a aucune importance pour la suite puisque les valeurs des variables sont des entiers pouvant atteindre 8000. Le tableau T2 donne les statistiques élémentaires correspondantes. Notons que les coefficients de corrélation entre populations ont augmenté.

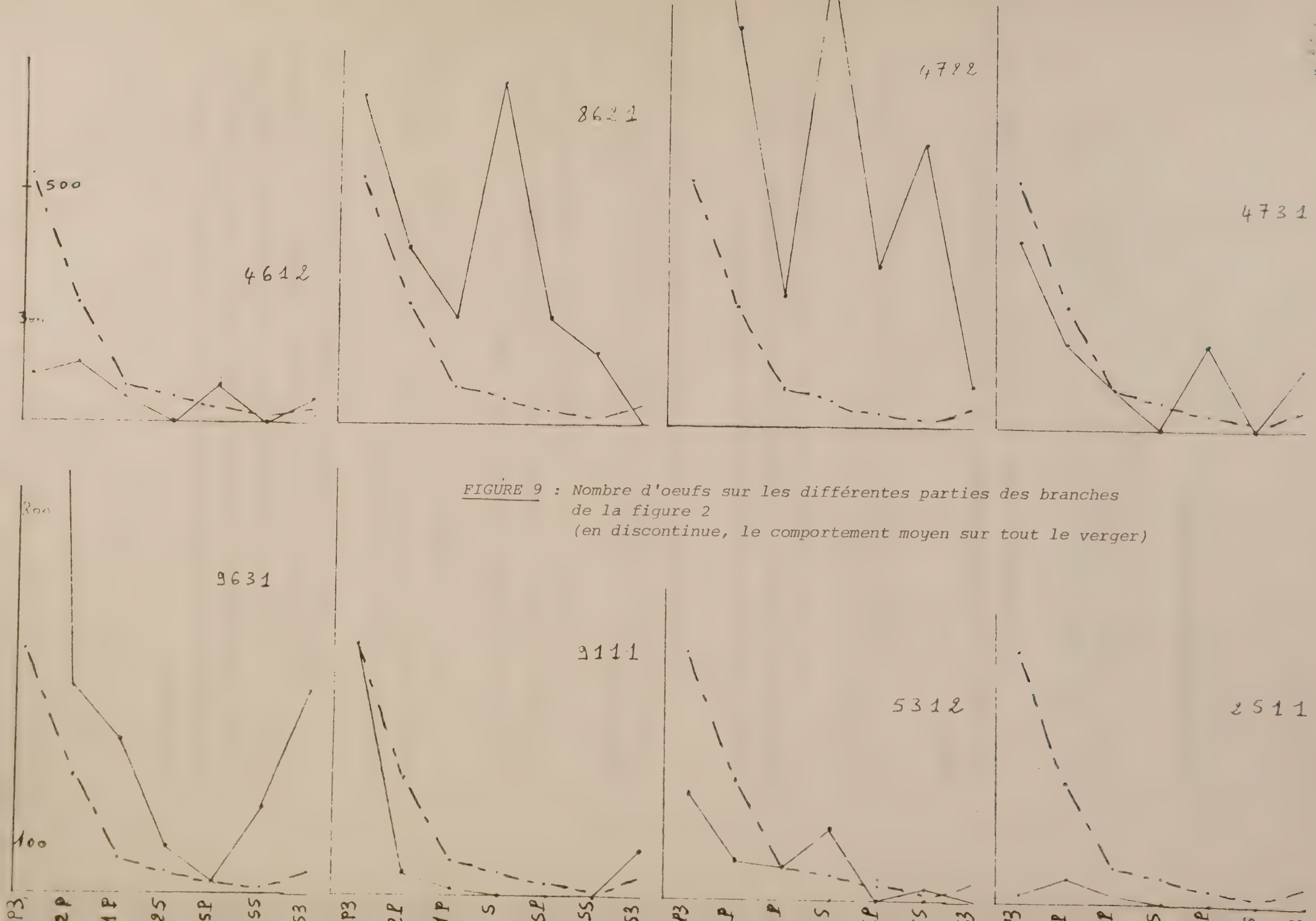


FIGURE 9 : *Nombre d'oeufs sur les différentes parties des branches de la figure 2*
(en discontinue, le comportement moyen sur tout le verger)

Le tableau T3 donne les coefficients de PEARSON de symétrie ($\sqrt{B1}$) et d'applatissement B2 utilisés pour tester la normalité des distributions observées ([1],p.66).

Dans notre cas, (378 observations), nous acceptons l'hypothèse de normalité si ([ ], p.207).

- au niveau 5%

$|\sqrt{B1}| < 0.3$ et $2.6 < B2 < 3.4$

- au niveau 1%

$|\sqrt{B1}| < 0.21$ et $2.5 < B2 < 3.7$

L'hypothèse de normalité est rejetée pour toutes les populations initiales. Les variables transformées ont toutes un comportement proche de la normale, cependant l'hypothèse de normalité n'est acceptée que pour LPTOT au niveau de 5% et pour LPOP3 au niveau de 1%. Cette hypothèse est rejetée pour les populations P2P et P2M à cause de la symétrie ($\sqrt{B1}$) pourtant proche des valeurs critiques correspondant au seuil 5%.

	VARIABLES INITIALES		VARIABLES TRANSFORMEES	
	$\sqrt{B_1}$	B_2	$\sqrt{B_1}$	B_2
PTOT	2.6	13.3	-0.27	2.74
POP3	3.8	26	-0.21	2.77
P2P	2	8.5	-0.32	2.86
P2M	2.2	9.4	-0.36	2.96
P1P	2	7.5	-0.69	3.40
P12PM	1.8	6.4	-0.5	2.90
P12M	1.6	5.4	-0.5	2.88
P1M	1.6	5.6	-0.47	2.91
P2S	4.6	30	0.84	2.01
P1SP	3	13.3	0.16	1.61
P1SS	7.9	88	1.4	3.6
P1S3	3.4	16.8	0.61	1.91

Tableau 3 : valeurs des coefficients de Pearson pour le test de l'hypothèse de normalité

STATISTIQUES CALCULEES SUR 378 INDIVIDUS. NOMBRE DE DEGRES DE LIBERTE= 377 SOMME DES POIDS= 378

	VARIABLE	MOYENNE	ECART-TYPE	VARIANCE	COEF.DE VARIA	MINIM...	MAXIMUM
1	PTOT	0.9730952E+03	0.9376489E+03	0.8791855E+06	0.9635736E+00	0.3500000E+02	0.7257000E+04
2	POP3	0.5202617E+03	0.5825747E+03	0.3393933E+06	0.1119772E+01	0.1400000E+02	0.5755000E+04
3	P2P	0.2548122E+03	0.2429625E+03	0.5903079E+05	0.9534965E+00	0.6000000E+01	0.1620000E+04
4	P2M	0.2234186E+03	0.2124335E+03	0.4512798E+05	0.9508314E+00	0.3500000E+01	0.1430000E+04
5	P1P	0.7465872E+02	0.8297443E+02	0.6884754E+04	0.1111382E+01	0.0000000E+00	0.4650000E+03
6	P12PM	0.3381322E+02	0.3506851E+02	0.1229800E+04	0.1037124E+01	0.0000000E+00	0.2090000E+03
7	P12M	0.3316011E+02	0.3354367E+02	0.1125178E+04	0.1011566E+01	0.0000000E+00	0.1715714E+03
8	P1M	0.3282616E+02	0.3250943E+02	0.1056863E+04	0.9903514E+00	0.0000000E+00	0.1601250E+03
9	P2S	0.4772221E+02	0.1171630E+03	0.1372717E+05	0.2455104E+01	0.0000000E+00	0.9990000E+03
10	P1SP	0.3369576E+02	0.5958038E+02	0.3549822E+04	0.1768187E+01	0.0000000E+00	0.4140000E+03
11	P1SS	0.1328836E+02	0.4465259E+02	0.1993854E+04	0.3360278E+01	0.0000000E+00	0.5980000E+03
12	P1S3	0.2705026E+02	0.5730615E+02	0.3283995E+04	0.2118505E+01	0.0000000E+00	0.4230000E+03
13	PBRS	0.1923808E+02	0.6989923E+02	0.4885902E+04	0.3633378E+01	0.0000000E+00	0.8450000E+03
14	PBRSM	0.1574603E+02	0.6401636E+02	0.4098094E+04	0.4065555E+01	0.0000000E+00	0.8450000E+03
15	LTOT	0.1925079E+03	0.7626007E+02	0.5815598E+04	0.3961399E+00	0.5900000E+02	0.4600000E+03
16	LONG3	0.2333862E+02	0.7457371E+01	0.5561237E+02	0.3195291E+00	0.9000000E+01	0.5200000E+02
17	L2P	0.2547618E+02	0.1123363E+02	0.1261945E+03	0.4409466E+00	0.6000000E+01	0.8100000E+02
18	L2M	0.2411182E+02	0.1073997E+02	0.1153471E+03	0.4454237E+00	0.9000000E+01	0.8100000E+02
19	L1P	0.5987300E+02	0.2277783E+02	0.5188296E+03	0.3804358E+00	0.2500000E+02	0.1340000E+03
20	L12PM	0.3002275E+02	0.1157687E+02	0.1340240E+03	0.3856033E+00	0.1250000E+02	0.8900000E+02
21	L12M	0.3046289E+02	0.1122334E+02	0.1259632E+03	0.3684264E+00	0.1300000E+02	0.7200000E+02
22	L1M	0.3038269E+02	0.1042063E+02	0.1085896E+03	0.3429792E+00	0.1300000E+02	0.6950000E+02
23	L2S	0.8211639E+01	0.1370476E+02	0.1878204E+03	0.1668942E+01	0.0000000E+00	0.8100000E+02
24	L1SP	0.3410052E+02	0.3563840E+02	0.1270095E+04	0.1045097E+01	0.0000000E+00	0.1600000E+03
25	L1SS	0.1443651E+02	0.2598558E+02	0.6752502E+03	0.1799991E+01	0.0000000E+00	0.1970000E+03
26	L1S3	0.2770898E+02	0.3875461E+02	0.1501919E+04	0.1398629E+01	0.0000000E+00	0.1940000E+03

MATRICE DE CORRELATIONS

		1 PTOT	2 POP3	3 P2P	4 P2M	5 P1P	6 P12PM	7 P12M	8 P1M	9 P2S	10 P1SP	11 P1SS	12 P1S3	13 PBRS
1	PTOT	1000												
2	POP3	948	1000											
3	P2P	775	601	1000										
4	P2M	660	491	929	1000									
5	P1P	611	450	606	580	1000								
6	P12PM	634	477	590	560	901	1000							
7	P12M	653	494	594	579	898	978	1000						
8	P1M	667	514	603	587	888	952	968	1000					
9	P2S	556	480	249	110	190	275	292	271	1000				
10	P1SP	494	330	475	419	440	554	572	526	263	1000			
11	P1SS	518	423	245	118	228	292	352	337	811	292	1000		
12	P1S3	437	482	165	180	264	249	271	393	-15	129	85	1000	
13	PBRS	493	598	182	124	109	164	175	150	365	2	223	180	1000
14	PBRSM	511	612	194	135	127	184	198	169	361	14	251	199	971
15	LTOT	458	447	199	112	186	203	211	223	351	402	330	423	128
16	LONG3	129	177	-40	-95	-18	47	51	31	104	95	108	126	58
17	L2P	160	139	349	355	-52	-72	-68	-36	-46	-48	-42	-15	40
18	L2M	79	55	292	382	-65	-85	-82	-46	-113	-76	-110	-10	-29
19	L1P	251	235	173	169	349	322	304	305	45	148	51	150	61
20	L12PM	287	277	172	153	251	389	351	340	124	250	104	164	184
21	L12M	246	246	116	109	201	310	316	296	110	234	116	158	153
22	L1M	228	224	104	95	217	304	302	336	72	191	89	256	63
23	L2S	231	221	8	-158	-23	-5	-25	-34	657	25	426	-123	204
24	L1SP	294	219	272	230	171	186	200	173	113	663	121	84	-17

Tableau 1 : Statistiques élémentaires sur les variables initiales

STATISTIQUES CALCULEES SUR 378 INDIVIDUS. NOMBRE DE DEGRES DE LIBERTE= 377 SOMME DES POIDS= 378

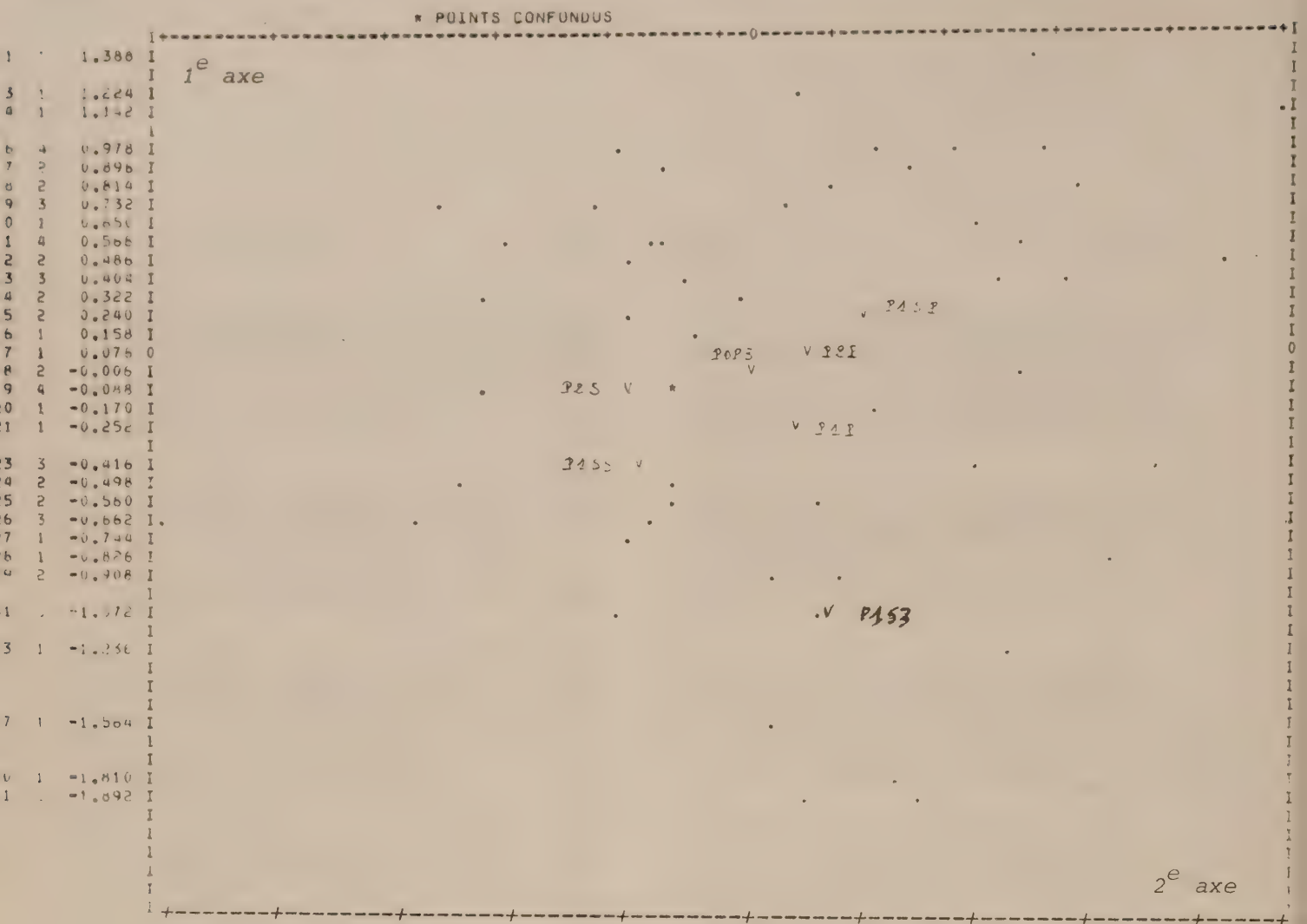
	VARIABLE	MOYENNE	ECART-TYPE	VARIANCE	COEF.DE VARIA	MINIMUM	MAXIMUM
1	PTOT	0.6484215E+01	0.9360890E+00	0.8762625E+00	0.1443642E+00	0.3526359E+01	0.8889856E+01
2	POP3	0.5795228E+01	0.1000407E+01	0.1000814E+01	0.1726260E+00	0.2708048E+01	0.8657995E+01
3	P2P	0.5129308E+01	0.9649230E+00	0.9310763E+00	0.1881195E+00	0.1945909E+01	0.7390796E+01
4	P2M	0.5005482E+01	0.9593067E+00	0.9202694E+00	0.1916512E+00	0.1504076E+01	0.7266127E+01
5	P1P	0.3686703E+01	0.1296288E+01	0.1680364E+01	0.3516119E+00	-0.4172325E-06	0.6144182E+01
6	P12PM	0.3012545E+01	0.1149957E+01	0.1322400E+01	0.3817227E+00	-0.4172325E-06	0.5347105E+01
7	P12M	0.3012550E+01	0.1126054E+01	0.1267997E+01	0.3737875E+00	-0.4172325E-06	0.5150809E+01
8	P1M	0.3038394E+01	0.1076734E+01	0.1159355E+01	0.3543758E+00	-0.4172325E-06	0.5082178E+01
9	P2S	0.1567786E+01	0.2183714E+01	0.4768606E+01	0.1392864E+01	-0.4172325E-06	0.6907753E+01
10	P1SP	0.2106495E+01	0.1872308E+01	0.3505536E+01	0.8888261E+00	-0.4172325E-06	0.6028276E+01
11	P1SS	0.9348564E+00	0.1560189E+01	0.2434190E+01	0.1668907E+01	-0.4172325E-06	0.6395260E+01
12	P1S3	0.1643474E+01	0.1862452E+01	0.3468725E+01	0.1133241E+01	-0.4172325E-06	0.6049730E+01
13	PBRS	0.8746603E+00	0.1689652E+01	0.2854925E+01	0.1931781E+01	-0.4172325E-06	0.6740517E+01
14	PBRSM	0.7928512E+00	0.1570613E+01	0.2466825E+01	0.1980968E+01	-0.4172325E-06	0.6740517E+01
15	LTOT	0.5188106E+01	0.3981491E+00	0.1585227E+00	0.7674265E-01	0.4094342E+01	0.6133395E+01
16	LONG3	0.3145541E+01	0.3068184E+00	0.9413749E-01	0.9754074E-01	0.2302584E+01	0.3970289E+01
17	L2P	0.3198129E+01	0.3894369E+00	0.1516611E+00	0.1217702E+00	0.1945909E+01	0.4406719E+01
18	L2M	0.3148311E+01	0.3759811E+00	0.1413618E+00	0.1194231E+00	0.2302584E+01	0.4406718E+01
19	L1P	0.4038232E+01	0.3802516E+00	0.1445913E+00	0.9416288E-01	0.3258095E+01	0.4905273E+01
20	L12PM	0.3370221E+01	0.3574365E+00	0.1277608E+00	0.1060573E+00	0.2602688E+01	0.4499808E+01
21	L12M	0.3387927E+01	0.3491141E+00	0.1218806E+00	0.1030465E+00	0.2639055E+01	0.4290458E+01
22	L1M	0.3393134E+01	0.3268135E+00	0.1068071E+00	0.9631610E-01	0.2639055E+01	0.4255611E+01
23	L2S	0.1097387E+01	0.1485636E+01	0.2207113E+01	0.1353793E+01	-0.4172325E-06	0.4406718E+01
24	L1SP	0.2533963E+01	0.1829644E+01	0.3347597E+01	0.7220484E+00	-0.4172325E-06	0.5081403E+01
25	L1SS	0.1193122E+01	0.1743498E+01	0.3039784E+01	0.1461290E+01	-0.4172325E-06	0.5288265E+01
26	L1S3	0.1964820E+01	0.1937794E+01	0.3755044E+01	0.9862449E+00	-0.4172325E-06	0.5272997E+01

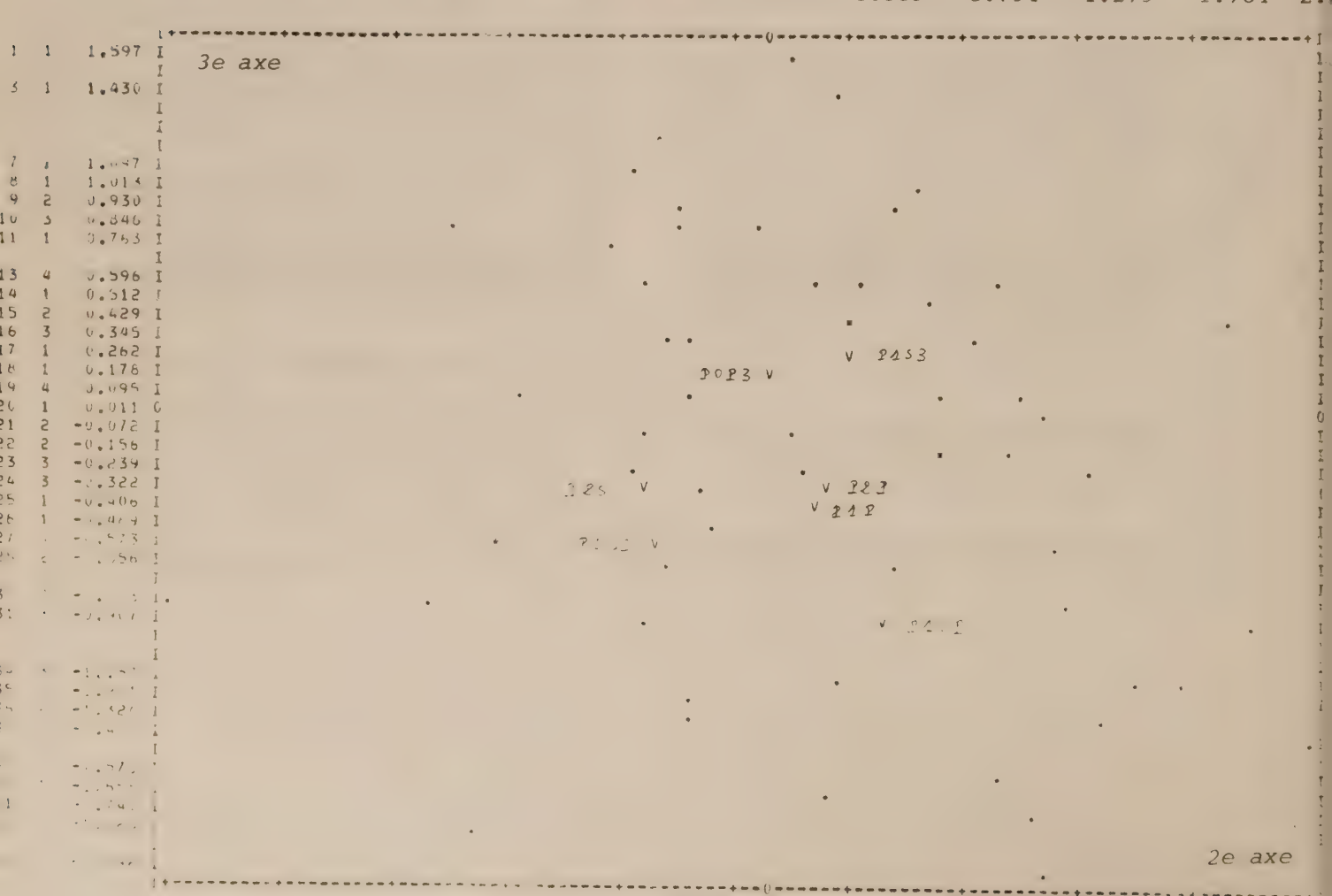
MATRICE DE CORRELATIONS

		1 PTOT	2 POP3	3 P2P	4 P2M	5 P1P	6 P12PM	7 P12M	8 P1M	9 P2S	10 P1SP	11 P1SS	12 P1S3	13 PBRS
1	PTOT	1000												
2	POP3	961	1000											
3	P2P	872	766	1000										
4	P2M	844	742	963	1000									
5	P1P	690	571	651	630	1000								
6	P12PM	736	620	681	665	961	1000							
7	P12M	746	628	686	675	944	986	1000						
8	P1M	767	655	690	682	914	958	970	1000					
9	P2S	279	249	79	-64	101	106	111	96	1000				
10	P1SP	491	397	442	422	442	497	507	485	106	1000			
11	P1SS	322	267	143	20	181	205	239	224	875	208	1000		
12	P1S3	295	327	95	127	166	198	210	293	-101	221	-14	1000	
13	PBRS	226	287	115	88	34	56	57	39	147	-127	47	102	1000
14	PBRSM	243	305	130	103	42	66	69	52	149	-119	54	113	994
15	LTOT	482	474	237	185	241	271	277	305	386	525	415	508	27
16	LONG3	95	143	-47	-88	-20	22	33	32	138	90	155	133	-53
17	L2P	176	182	320	305	-88	-94	-92	-81	-62	-141	-77	-123	72
18	L2M	115	122	264	323	-113	-119	-121	-105	-222	-179	-216	-109	5
19	L1P	339	328	243	249	378	350	334	333	34	181	49	178	30
20	L12PM	353	348	225	227	317	379	363	360	79	244	111	223	85
21	L12M	322	320	188	174	262	326	342	328	128	227	182	207	87
22	L1M	317	316	162	157	266	319	329	360	106	193	143	295	31
23	L2S	125	108	-59	-220	-6	-15	-15	-34	950	20	790	-146	139
24	L1SP	259	189	236	217	199	199	206	187	73	855	132	146	-150

Tableau 2 : Statistiques élémentaires sur variables transformées par $\alpha \rightarrow \log(X + 1)$

RAPHIQUE 1 : *Projection sur les plans définis par les 3 premiers axes de l'analyse des correspondances*

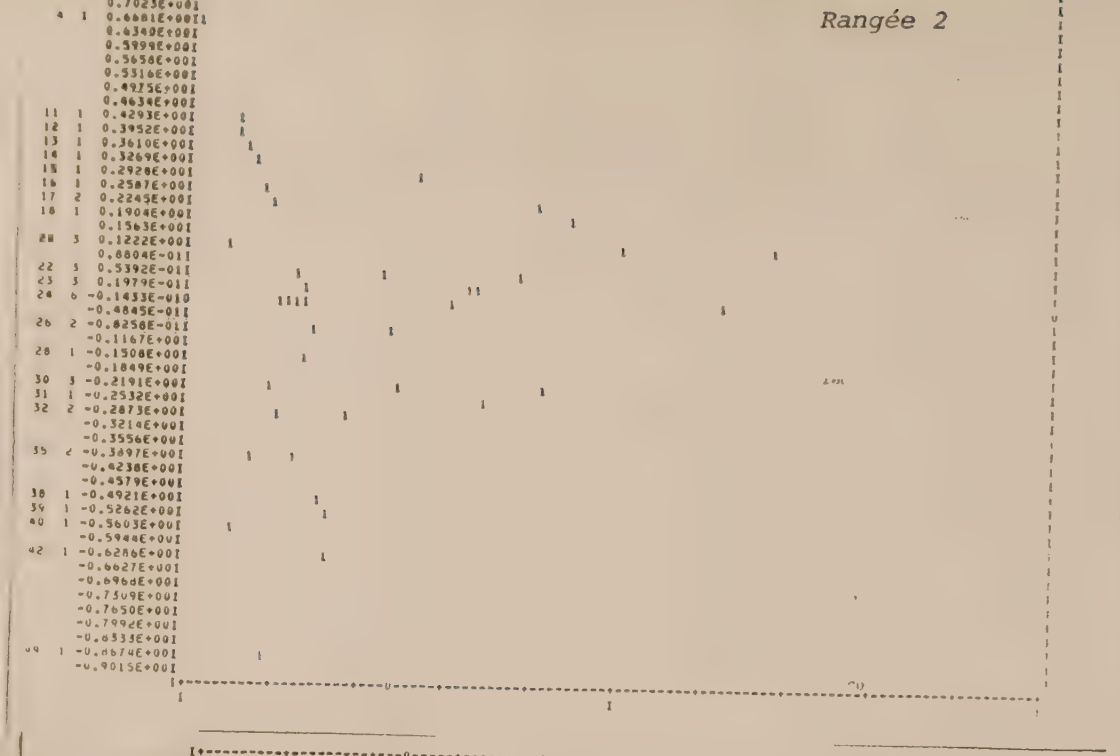


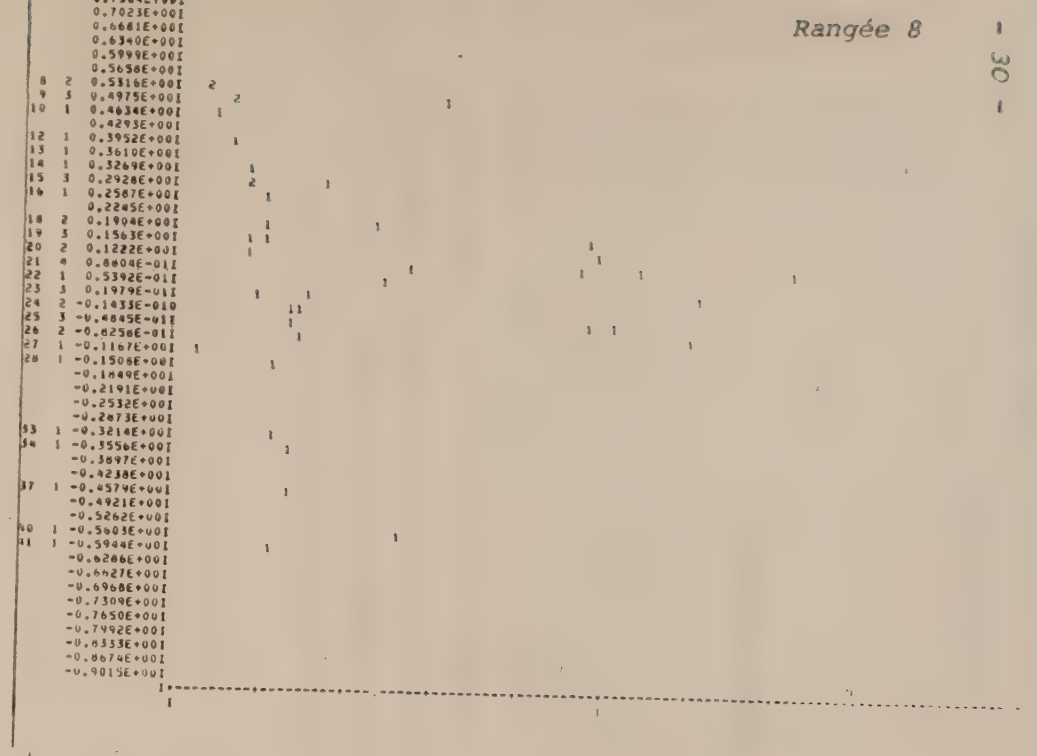


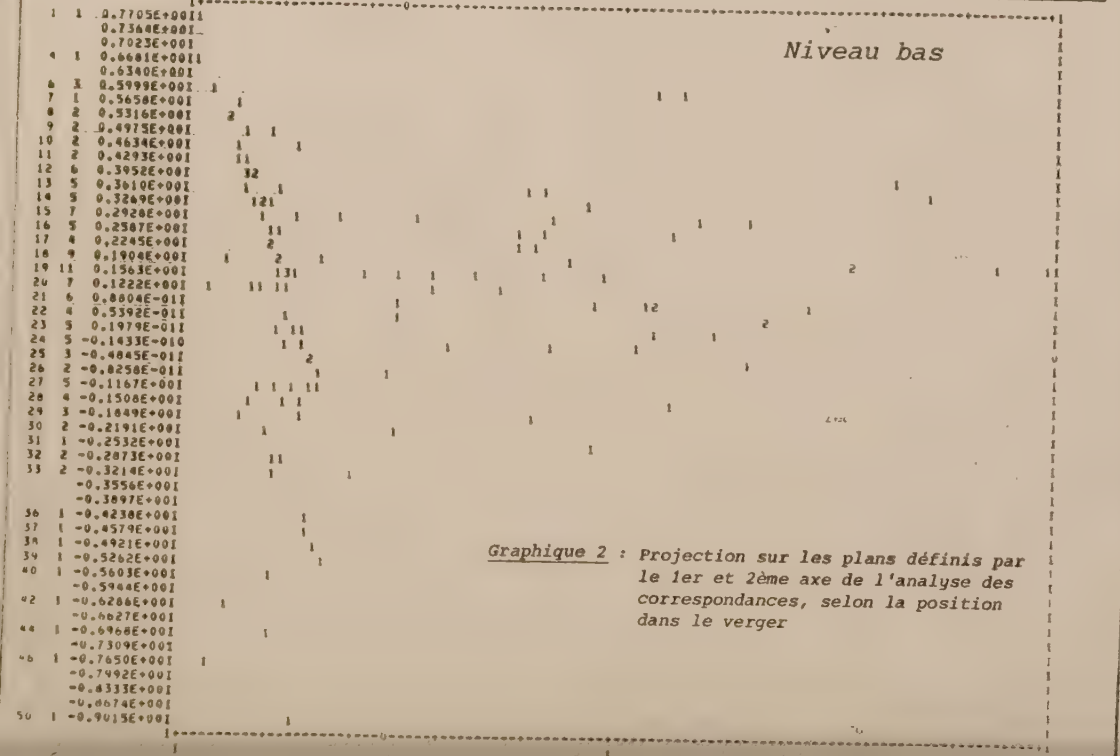
A.II-4) ANALYSE FACTORIELLE DES CORRESPONDANCES

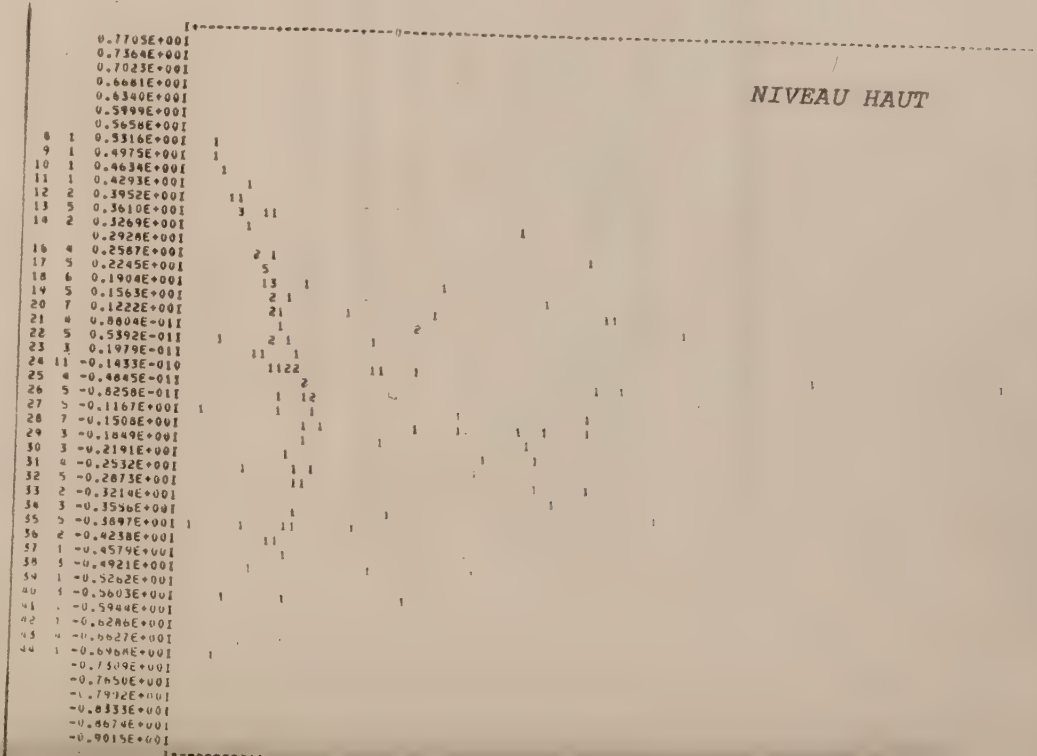
Comme première approche, nous avons d'abord cherché à dépouiller le tableau de nos données en étudiant le profil de répartition des oeufs sur les différentes parties de la branche de 3 ans. L'analyse des correspondances ([1],p.133) appliquée au tableau des comptages POP3, P2P, P2S, P1P, P1SP, P1SS et P1S3, indique notamment l'influence de la proximité de deux parties de la branche sur les profils des populations correspondantes. Cependant, l'influence des données manquantes (populations et longueurs nulles) est lourde dans la détermination des axes principaux de l'analyse.

En refaisant l'analyse sur seulement les 56 branches telles que L1SS et L1S3 non nulles, la même relation entre projections des variables subsiste (G1). D'autre part, les projections des branches par rangée, travée, niveau ou face sont semblables et laissent supposer que les relations entre populations ne varient pas avec la position de la branche dans le verger (G2-G3).



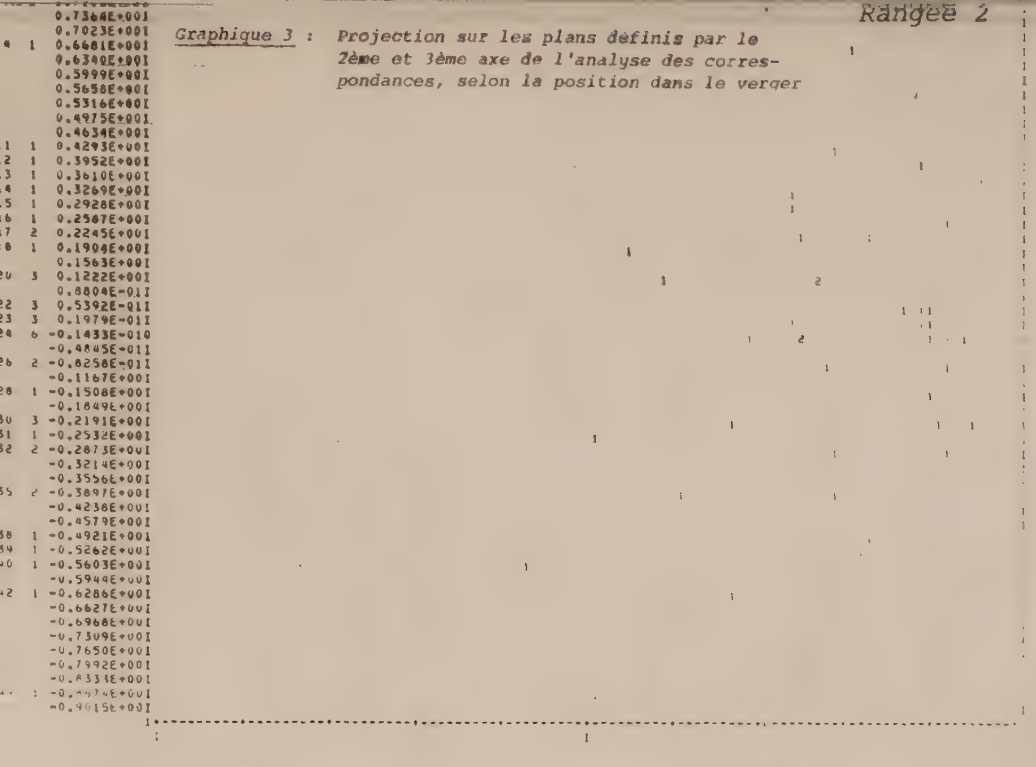


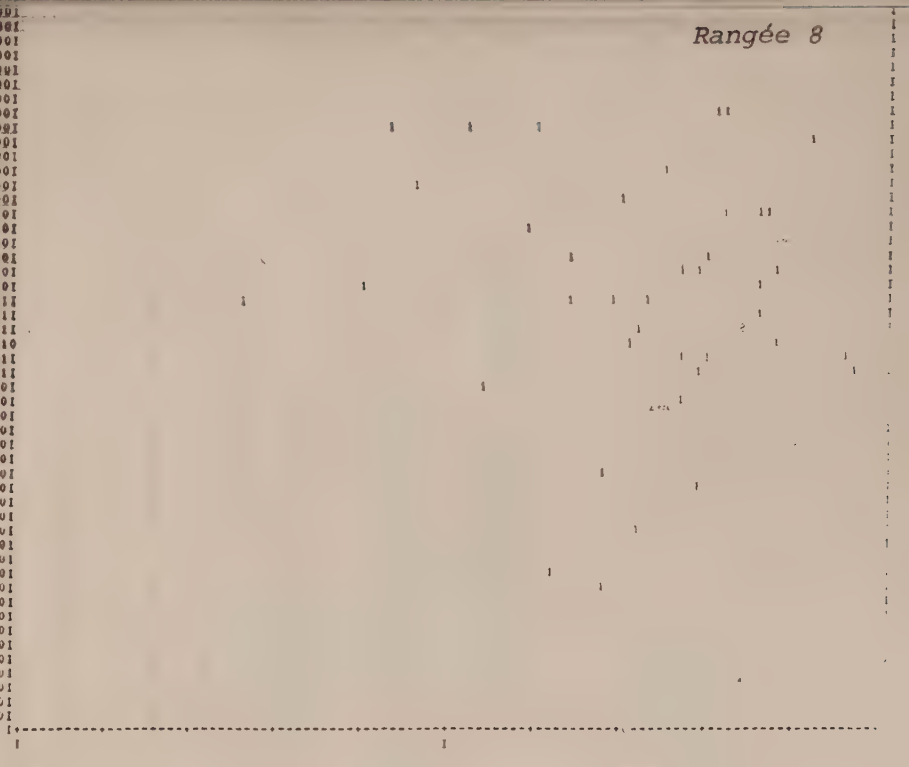


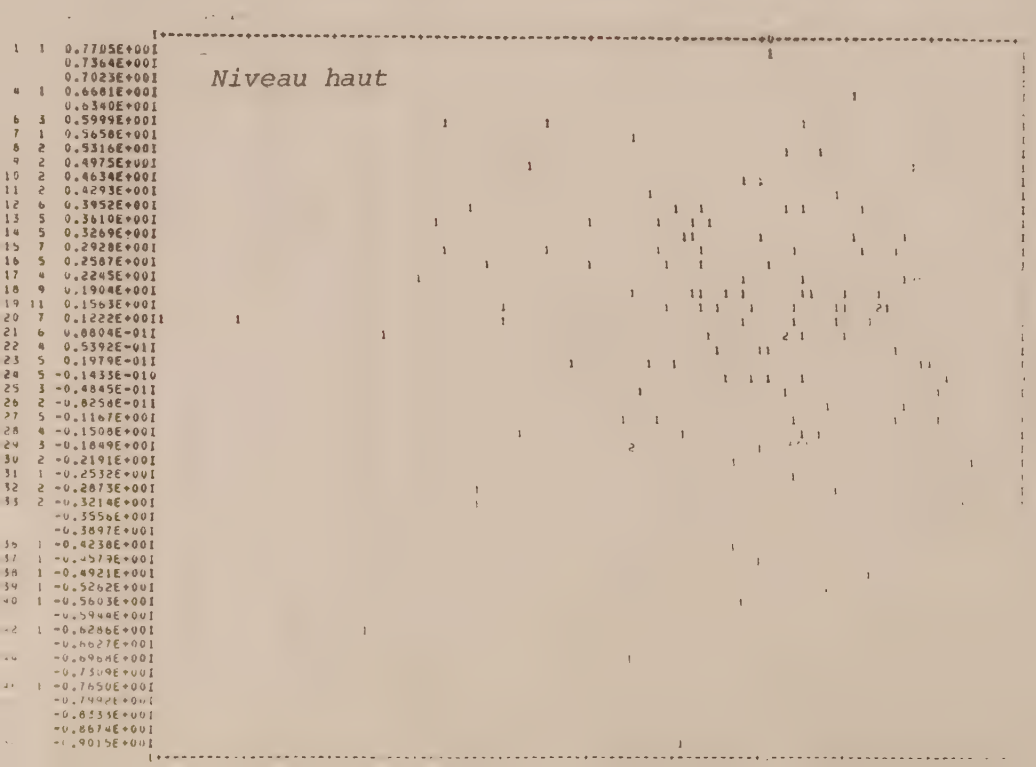


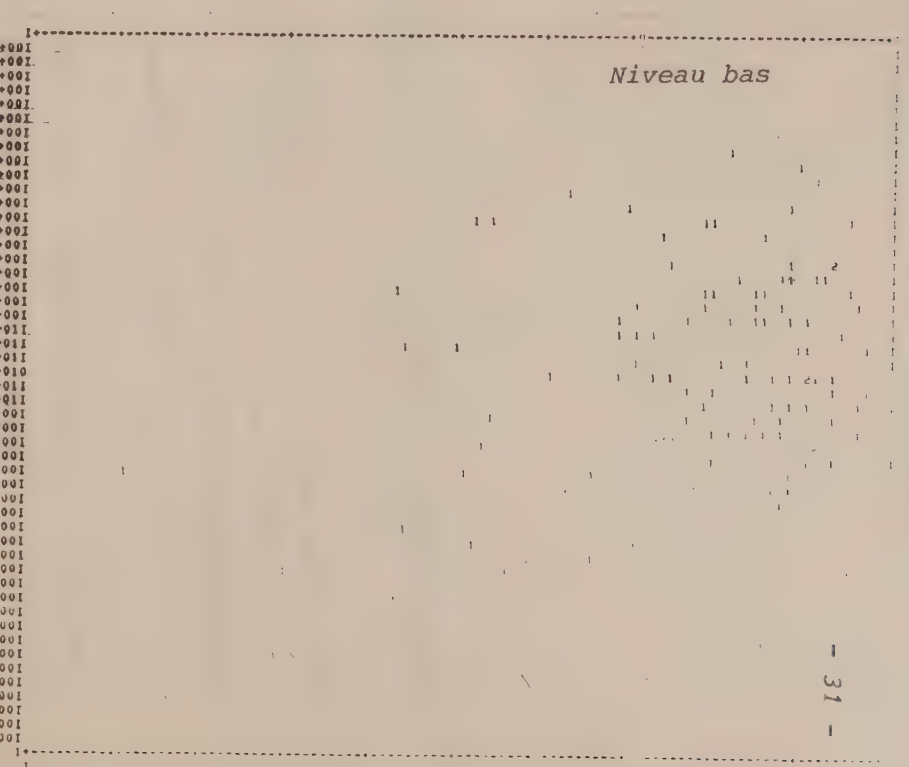
Graphique 2 : *Projection sur les plans définis par le 1er et 2ème axe de l'analyse des correspondances, selon la position dans le verger*

Graphique 3 : *Projection sur les plans définis par le 2ème et 3ème axe de l'analyse des correspondances, selon la position dans le verger*









A - II - 5) REGRESSIONS

a) Relations entre populations et caractéristiques architecturales de la branche.

Nous avons considéré plusieurs possibilités de liaison entre d'une part les populations principales et secondaires, totales ou moyennes, et d'autre part l'ensemble de toutes les caractéristiques architecturales de la branche (longueur, inclinaison, orientation cardinale...).

La transformation X → Log(X+1) est apparue nécessaire à l'examen des résidus : l'écart-type résiduel est proportionnel à la valeur estimée. Cette transformation a été appliquée soit seulement à la population, soit aux caractéristiques seules, ou encore aux deux (soit donc 24 variables à expliquer en fonction de 32). ([1], p. 209)

Les seules relations retenues sont les suivantes :

		coefficients	E.T.	R^2
LPOP3	=	-0.38	0.6	0.22
		+ 1.2 LLTOT	0.1	
LP2P	=	0.1	0.6	0.18
		+1.06 LL2PT	0.1	
LP1P	=	-1.5	0.6	0.14
		+1.3 LL1P	0.16	

La colonne E.T. indique l'écart-type des coefficients de la régression donnés sur la même ligne. R est le coefficient de corrélation multiple

Ces relations ont été retenues à cause des deux faits suivants :

- les coefficients des régressions sur les longueurs sont tous proches de la valeur 1, avec un écart-type très faible (de l'ordre de 0.1) -

S. TOTALE	80.054
S. MODELE	56.922
S. RESIDUELLE	23.132

ECART-TYPE RESIDUEL	0.525
COEFFICIENT DE CORRELATION MULTIPLE	0.843
TEST GLOBAL D'AJUSTEMENT F(2, 84)	103.351 **

EQUATION DE LA REGRESSION

COEFFICIENT	NO. VARIABLE	ECART-TYPE	T. STUDENT	BETA	VIF
-0.2367E-02	CONST	0.7195E+00	-0.0033		
0.6214E+00	X 39	0.5133E-01	12.1076 **	0.7250	1.0423
0.7266E+00	X 41	0.1404E+00	5.1557 **	0.3087	1.0423

NORME DU VECTEUR DE REGRESSION : 0.7880
(DANS L'ESPACE DES PARAMETRES)

TABLEAU 4 : L POP3 = -0.002 + 0.62 LPBRS + 0.73 LLTOT

CARTE GRAPHIQUE

POP = LONG * POP / LONG * LOG = ET CARACTERISTIQUES DES BRANCHES

ABSCISSE = Y ESTIME ORDONNEE = RESIDU SYMBOLES . POINTS
* POINTS SUPPLEMENTAIRES
0 POINTS CONFONDUS

LIGNE NO			
1	1	1.512	
6	2	1.162	
9	2	0.951	
12	2	0.741	
13	1	0.671	
14	1	0.601	
15	2	0.531	
16	1	0.461	
17	3	0.390	
18	2	0.320	
19	7	0.250	
20	6	0.180	
21	5	0.110	
22	3	0.040	0
23	6	-0.030	0
24	5	-0.100	.
25	3	-0.170	
26	5	-0.240	
27	5	-0.311	
28	5	-0.381	
30	5	-0.521	.
31	1	-0.591	
32	2	-0.661	
33	4	-0.731	
34	2	-0.801	
35	1	-0.871	
37	1	-1.012	
39	1	-1.152	
40	1	-1.222	

4.533 4.901 5.269 5.637 6.005 6.373 6.741 7.109 7.476 7.844 8.212

GRAPHIQUE 4 : LPOP3 = -0.002 + 0.62 LPBRS + 0.73 LLTOT

l'hypothèse de nullité de chacun de ces paramètres est rejetée au seuil de 1%. Les régressions étant faites sur les transformations logarithmiques, ceci laisse prévoir une proportionnalité entre longueurs et populations.

- La population sur le bois de 2 ans (resp. 3 ans) est liée à la somme des longueurs de tous les rameaux de 1 an et de celle du bois de 2 ans considéré (resp. tous les rameaux de 1 et 2 ans ainsi que de celle du bois de 3 ans considéré). Si nous ne considérons que la longueur du bois, alors :

		coefficient	E.T.	R^2
LPOP3	=	4.3	0.5	0.02
		+0.47 LLONG3	0.16	
LP2P	=	2.6	0.4	0.10
		+0.80 LL2P	0.1	

Remarquons que la longueur du bois de 3 ans se distingue des autres. Le bois étant plus rugueux et comportant plus d'obstables, l'interprétation de son rôle est moins évidente.

Des relations plus compliquées peuvent être données, par exemple :

LP2P	=	-0.25	0.6	0.18
		+0.9 LL2P	0.1	
		+0.75 LL12PT	0.1	

Dans tous les cas, les coefficients de corrélations R sont faibles. Ce qui s'explique par la structure des haies qui est supposée uniforme sur tout le verger, alors que la répartition des oeufs est plutôt aggrégative. L'architecture de la branche ne peut, à elle seule, expliquer l'effectif d'oeufs d'hiver de l'araignée rouge. Nous reprendrons cette question plus loin, en tenant compte de la position dans le verger.

b) Populations sur bourses

Si nous ne considérons comme variables explicatives que les populations sur bourses (PBRS, PBRSM, LPBRS), alors nous obtenons les 2 résultats suivants :

S. TOTALE	67.770
S. MODELE	45.557
S. RESIDUELLE	22.213

ECART-TYPE RESIDUEL	0.514
COEFFICIENT DE CORRELATION MULTIPLE	0.820
TEST GLOBAL D'AJUSTEMENT F(2, 84)	86.141 **

EQUATION DE LA REGRESSION

COEFFICIENT	NO. VARIABLE	ECART-TYPE	T. STUDENT	BETA	VIF
0.6914E+00	CONST	0.7051E+00	1.2642		
0.5397E+00	X 39	0.5030E-01	10.7315 **	0.6844	1.0425
0.7238E+00	X 41	0.1381E+00	5.2415 **	0.3343	1.0425

NORME DU VECTEUR DE REGRESSION : 0.7616
(DANS L'ESPACE DES PARAMETRES)

ATRICE DE CORRELATION DES ESTIMATEURS

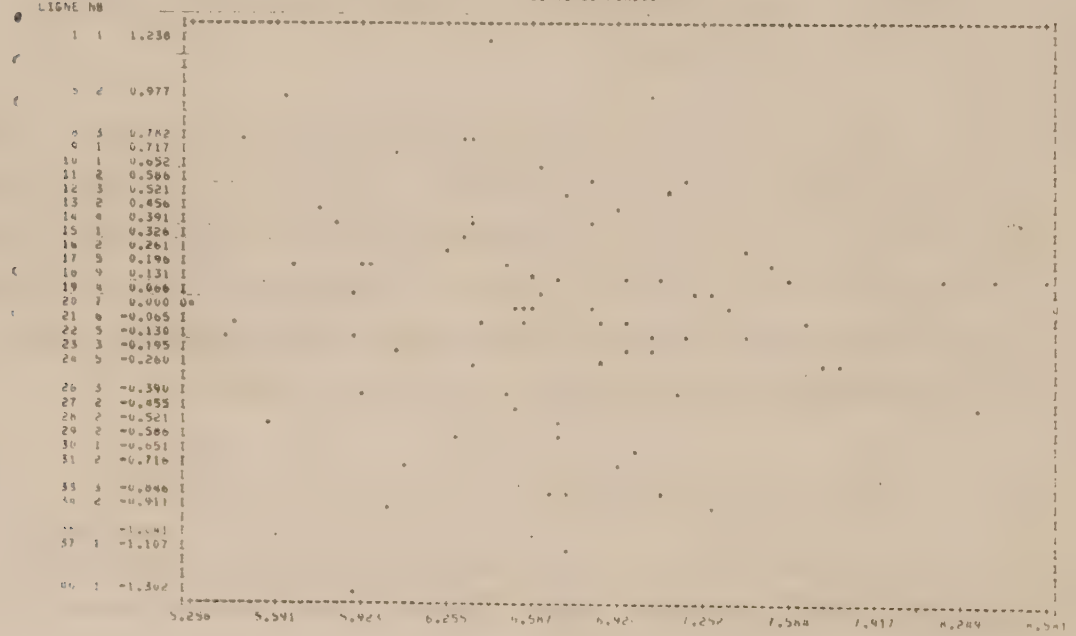
	39	41
39	1000	-201
41	-201	1000

TABLEAU 5 : LPTOT = -0.89 + 0.54 LPBRS + 0.72 LLTOT

CARTE GRAPHIQUE

PUP , LUGG , PUP / LONG , LUG , ET CARACTERISTIQUES DES BRANCHES

ABSCISSE = Y ESTIME ORDONNEE = RESIDU SYMBOLES . POINTS
* POINTS SUPPLEMENTAIRES
@ POINTS CONFONDUS



GRAPHIQUE 5 : LPTOT = 0.89 + 0.54 LPBRS + 0.72 LLTOT

		E.T	R^2
LPOP3 =	3.56	0.23	0.62
	+ 0.67 LPBRS	0.06	
LPTOT =	4.44	0.22	0.57
	+ 0.59 LPBRS	0.06	

L'analyse n'est faite que sur les 87 branches comportant des bourses sur le bois de 3 ans.

Une amélioration sensible est apportée par la longueur totale LLTOT parmi toutes les autres caractéristiques architecturales de la branche. Les 2 relations deviennent alors :

		E.T	R^2
LPOP3 =	0.	0.7	0.70
	+ 0.62 LPBRS	0.05	
	+ 0.72 LLTOT	0.14	
LPTOT =	0.89	0.71	0.67
	+ 0.54 LPBRS	0.05	
	+ 0.72 LLTOT	0.14	

(voir tableaux des résultats complets T4-T5 et graphes des résidus en fonction de l'estimée G4-G5).

Ce résultat -le fort pouvoir explicatif de la population sur bourse- confirme celui de GASSER et KLAY dans le cas où les caches désignent des bourses.

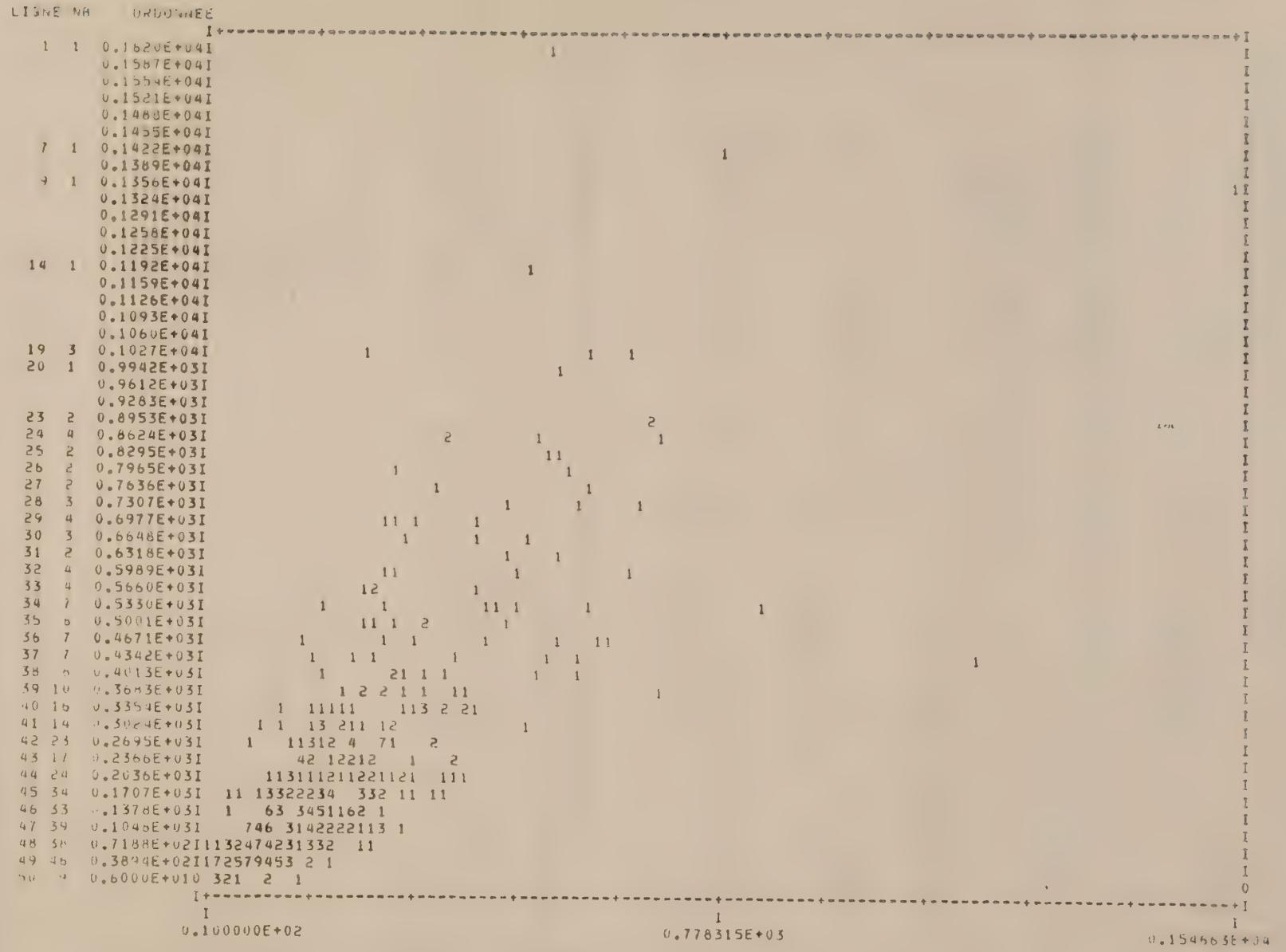
C) Relation entre POP2 et POP1 :

L'étude des relations entre populations donne de meilleurs résultats. Pour la population sur le bois de 2 ans P2P, nous avons retenu la relation suivante :

		E.T	R^2
LP2P =	3.2	0.1	0.47
	+ 0.49 LP12PT	0.03	

Tous les coefficients ayant un écart-type très faible et sont significativement non nuls au seuil de 1%. Si nous tenons compte des caractéristiques de la branche, nous obtenons :

		E.T	R^2
LP2P =	- 0.07	0.28	0.64
	+ 0.52 LP12PT	0.02	
	+ 0.97 LL2P	0.08	



GRAPHIQUE 7 : *Graphe de P2P en fonction de L2P* $\times \sqrt{P12PT}$

(tableau T6 et graphe des résidus en fonction des estimées G6).

L'hypothèse que la relation s'écrive :

$$LP2P = 0.5\ LP12T + LL2P$$

est acceptée au seuil de 1% ($F = 0.953 < 3.78$)

Notons α un bruit blanc $N(0,\sigma^2)$, où σ^2 est la variance résiduelle donnée dans T6, et soit $\beta = \exp \alpha$. β est de moyenne $\exp \frac{1}{2}\sigma^2$ et nous pouvons écrire :

$$P2P = \beta\ .\ L2P\ .\ \sqrt{P12PT}$$

Cette relation ne peut être dûe simplement à des liaisons autre longueurs correspondants. En effet la régression de la longueur du rameau de 2 ans sur les rameaux de 1 an donne des résultats où seule la constante est significativement non nulle et $R^2 = 0.02$

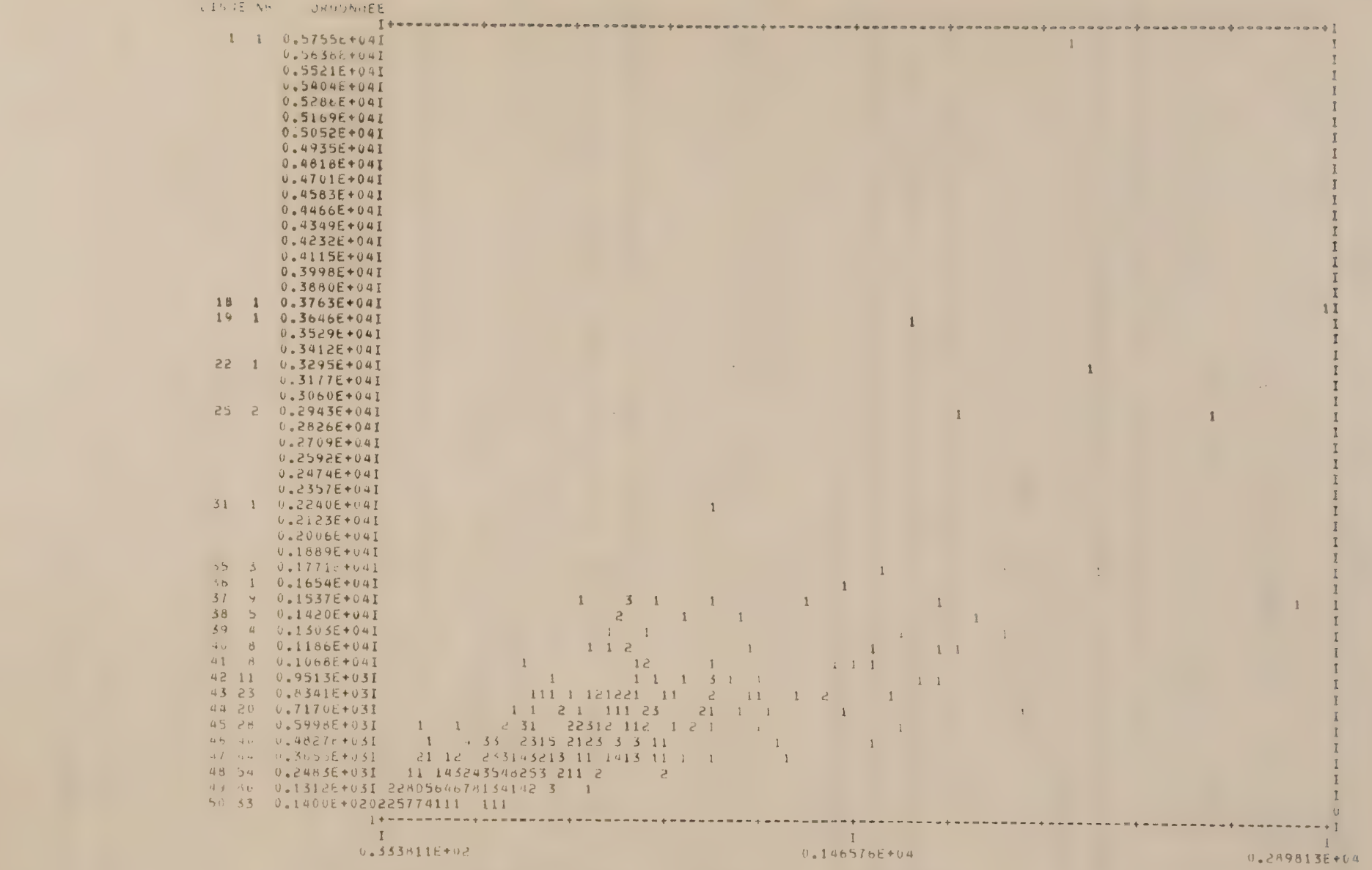
La bibliographie (I,[3]) nous permet de dire que l'araignée rouge pond de préférence sur la face inférieure des rameaux et sur les obstacles tels que cicatrices, rugosités, ect...

Si nous supposons que le nombre de tels obstacles est proportionnel à la longueur, alors $\frac{P2P}{L2P}$ s'interprète comme la densité d'oeufs déposés sur le rameau de 2 ans. La relation précédente signifie alors que cette densité est proportionnelle à la racine carrée du total de la population sur les bois de 1 an portés par le rameau considéré.

d) Relation entre POP3 et les populations plus jeunes. Si nous ne considérons que les populations d'âge inférieur comme variables explicatives, la meilleure relation obtenue lie la population sur le bois de 3 ans aux populations P1T et P2T.

		E.T	R^2
LPOP3 =	-0.34	0.33	0.70
	+0.85 Log(P2T + P1T)	0.03	
	+0.4 LLONG3	0.09	

(voir tableau T8 pour plus de précision ainsi que le graphe G8 des résidus en fonction de la valeur estimée).



NB. DE DONNEES MANQUANTES EN X SEUL= 0 EN X ET Y = 0 EN Y SEUL= 0

GRAPHIQUE 9 : Graphe des POP3 en fonction de $\sqrt{LONG3} = (P1T+P2T)^{4/5}$

Des relations aussi bonnes peuvent être obtenues telles que :

			E.T	R^2
LPOP3	=	-0.5	0.4	0.70
		+0.65 LP2T	0.04	
		+0.13 LP1T	0.03	
		+0.45 LLTOT	0.08	
	=	0.15	0.3	0.69
		+0.66 LP2T	0.04	
		+0.18 LP1T	0.03	
		+0.43 LLONG3	0.09	

Un fait remarquable dans toutes ces relations est que tous les coefficients, et surtout ceux des populations, sont déterminés avec un très faible écart-type.

Nous écrivons, comme relation approximative :

$$POP3 = \beta . \sqrt{LONG3} . (P1T + P2T)^{4/5}$$

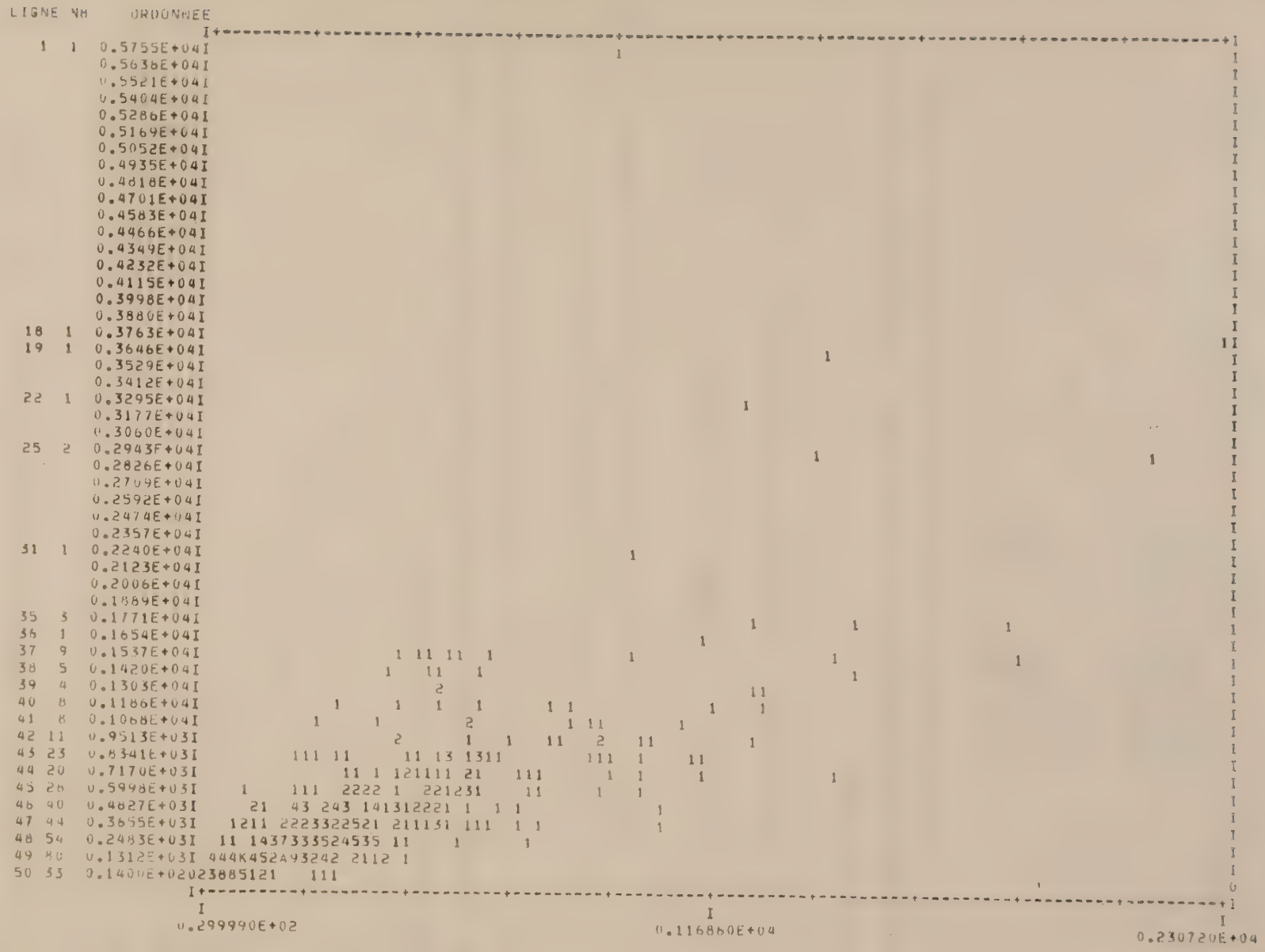
($F=15 > F_{\alpha},3,375 = 2,60$ pour $\alpha = 5\%$), où β est de moyenne $exp(\frac{1}{2}\sigma^2)$, σ^2 étant l'écart type résiduel de la première relation (G9).

La comparaison avec l'expression de P2P en fonction de P12T n'est pas évidente. Si nous ne considérons pas les caractéristiques de la branche comme variables explicatives alors

			E.T.	R^2
LPOP3	=	0.91	0.17	0.68
		+0.86 Log(P1T+P2T)	0.03	

Les coefficients étant significativement non nuls au seuil de 1%. La relation entre POP3 et (P1T+P2T) est donc plus proche de la linéarité que celle liant P2P à P12PT. (La comparaison avec l'expression précédente montre que l'apport des longueurs est ici très faible).

Si nous ne tenons compte que des relations entre la population sur le bois de 3 ans et celles sur les bois de 2 ans, alors nous obtenons :



GRAPHIQUE 11 : Graphe de POP3 en fonction de $(P2T)^{4/5} . \sqrt{LONG3}$

		E.T	R^2
LPOP3 =	- 1.2	0.4	0.69
	+ 0.75 LP2T	0.03	
	+ 0.6 LLTOT	0.08	
=	- 0.2	0.35	0.65
	+ 0.8 LP2T	0.03	
	+ 0.5 LLONG3	0.10	

(voir T10, G10)

tous les paramètres étant significativement non nuls au niveau de 1%, sauf la constante dans la deuxième expression.

L'hypothèse que LPOP3 s'écrive :

$$LPOP3 = 0.8\ LP2T + 0.5\ LLONG3$$

est acceptée au niveau de 1%, et nous pouvons écrire :

$$POP3 = \beta \cdot \sqrt{LONG3} \cdot (P2T)^{4/5} \quad (F = 0.09 < 3.78)$$

qui est similaire à l'expression de POP3 en fonction de (P1T+P2T) et Long3 (voir G11).

L'apport des longueurs est faible par rapport à la simple équation :

LPOP3 =	1.44	0.17	0.64
	+0.82 LP2T	0.03	

Si par contre, nous ne tenons pas compte de P2T, alors :

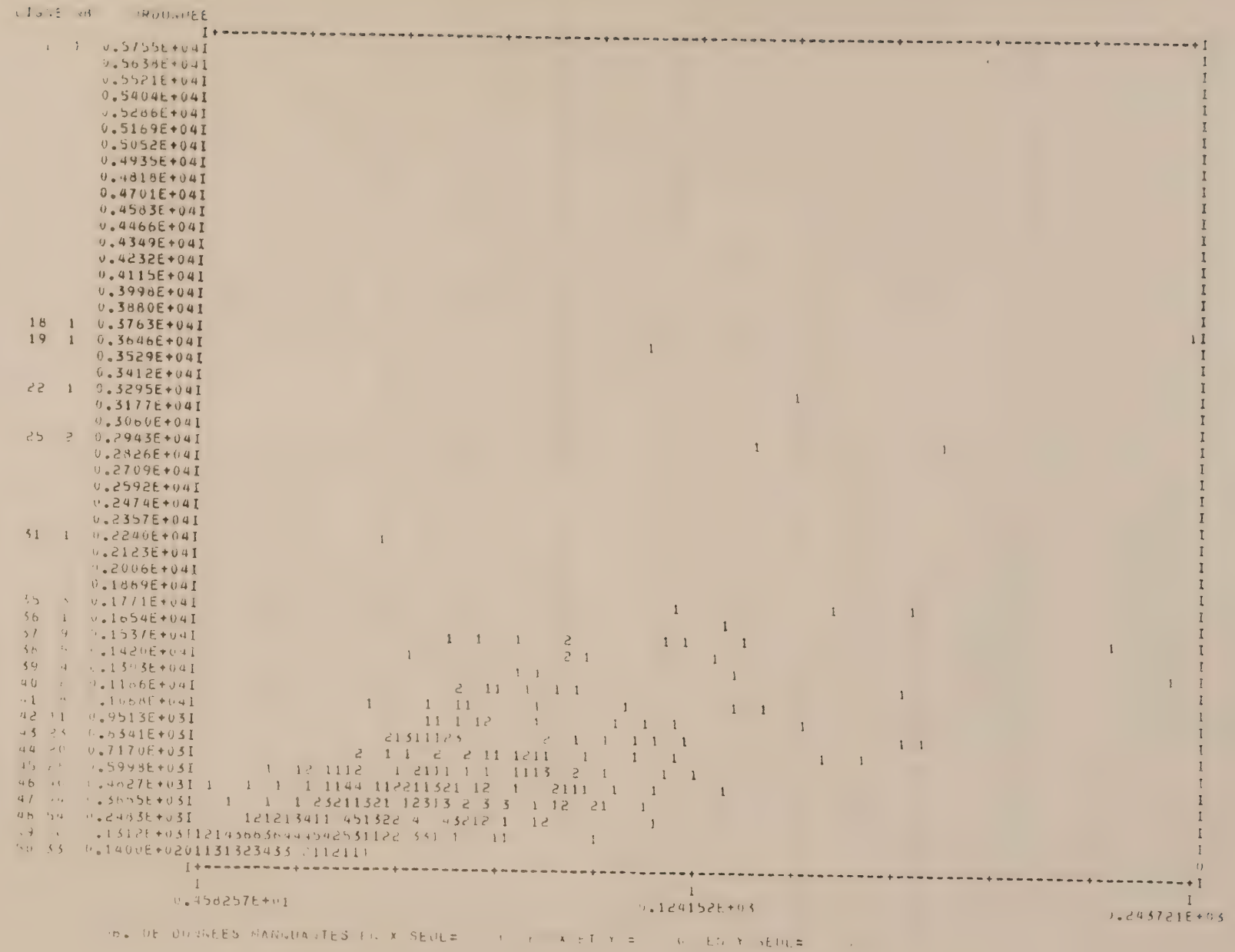
		E.T	R^2
LPOP3 =	1.7	0.28	0.53
	+0.53 LP1T	0.03	
	+0.53 LL2T	0.07	
=	1.6	0.5	0.48
	+0.45 LP1T	0.03	
	+0.43 LLTOT	0.1	
=	0.5	0.47	0.54
	+0.54 LP1T	0.03	
	+0.23 LLONG3	0.12	
	+0.69 LL2P	0.09	

(voir T12, G12)

La première relation parait mieux convenir, et nous l'écrirons d'une façon approximative (G13)

$$POP3 = \beta \sqrt{P1T}\ \sqrt{L2T} \quad \text{(hypothèse rejetée à 5\%)}$$

qui nous rapporche de l'expression de P2P en fonction de P12T, mais qui



GRAPHIQUE 13 : Graphe de POP3 en fonction de $\sqrt{P1T}$ x $\sqrt{P2T}$

est moins bonne que l'expression de POP3 quand nous tenons compte de (P2T + P1T) ou seulement de P2T.

Là aussi, l'apport des longueurs est négligeable. Nous avons :

$$LPOP3 = \underset{0.03}{\overset{3.54}{+0.52\ LP1T}} \qquad 0.13 \qquad 0.48$$

S. TOTALE	351.013
S. MODELE	221.950
S. RESIDUELLE	129.065

ECART-TYPE RESIDUEL	0.587
COEFFICIENT DE CORRELATION MULTIPLE	0.795
TEST GLOBAL D'AJUSTEMENT F(2, 375)	322.439 **

EQUATION DE LA REGRESSION

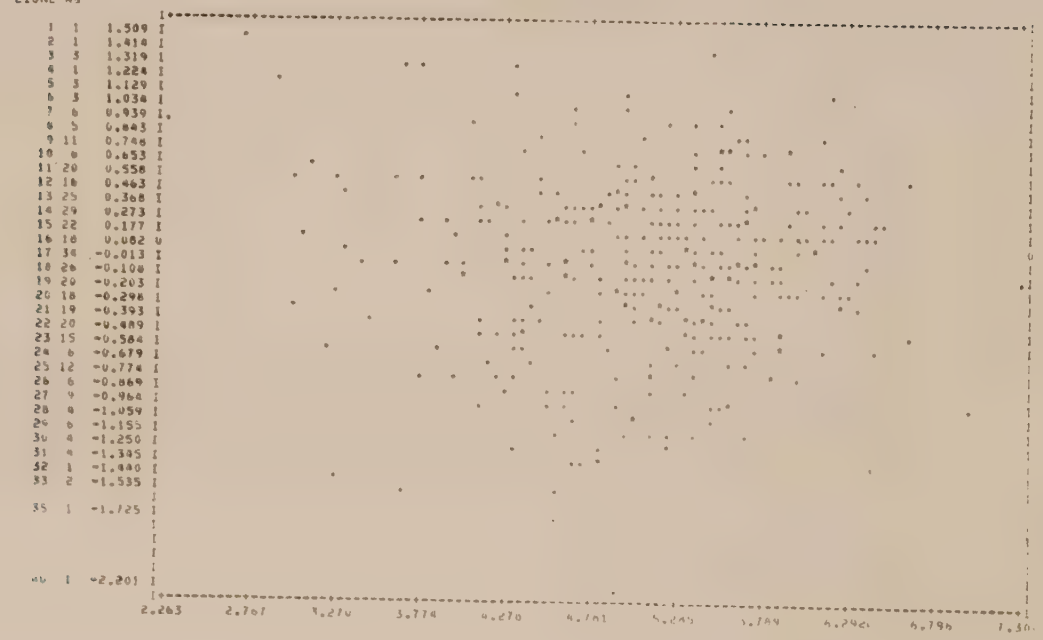
COEFFICIENT	NO. VARIABLE	ECART-TYPE	T. STUDENT	BETA	VIF
-0.7112E-01	CONST	0.2750E+00	-0.2586		
0.5192E+00	X 83	0.2234E-01	23.2449 **	0.7315	1.0100
0.9734E+00	X 43	0.7797E-01	12.4836 **	0.3928	1.0100

TABLEAU 6 : LP2P = -0.07 + 0.52 LP12PT + 0.97 LL2P

CARTE GRAPHIQUE

POP , LONG , POP / LONG , LOG , ET CARACTERISTIQUES DES BRANCHES

ABSCISSE = Y ESTIME ORDONNEE = RESIDU SYMBOLES . POINTS
+ POINTS SUPPLEMENTAIRES
* POINTS CONFONDUS



GRAPHIQUE 6 : LP2P = -0.07 + 0.52 LP12PT + 0.97 LL2P

S. TOTALE	377.304
S. MODELE	265.898
S. RESIDUELLE	111.406

ECART-TYPE RESIDUEL	0.545
COEFFICIENT DE CORRELATION MULTIPLE	0.839
TEST GLOBAL D'AJUSTEMENT F(2, 375)	447.514 **

EQUATION DE LA REGRESSION

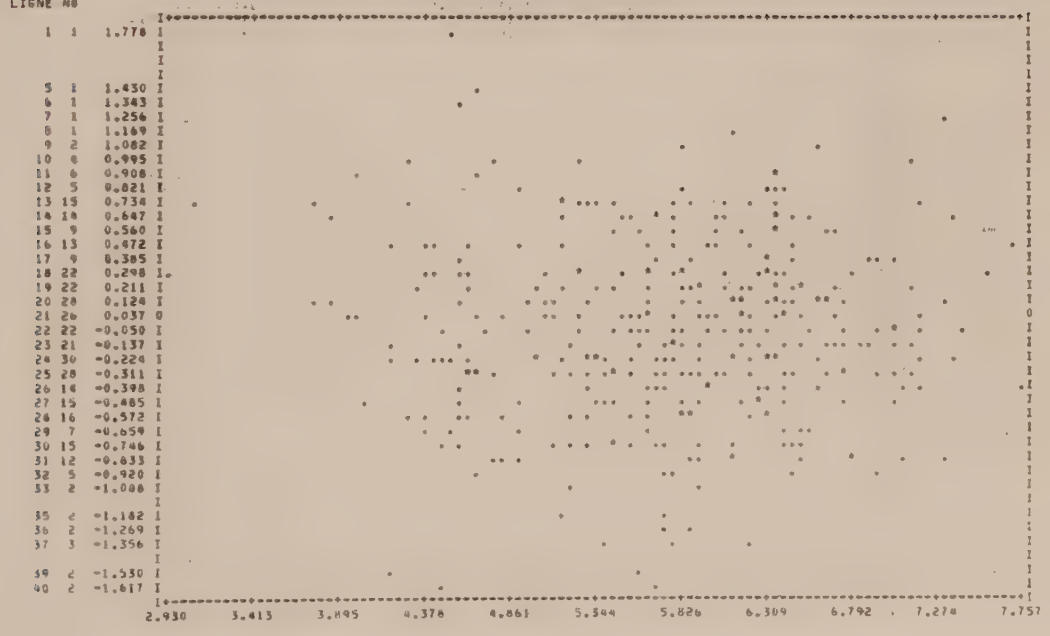
COEFFICIENT	NO.	VARIABLE	ECART-TYPE	T. STUDENT	BETA	VIF
-0.3411E+00		CONST	0.3297E+00	-1.0345		
0.8535E+00		X 84	0.2895E-01	29.4793 **	0.8274	1.0006
0.4029E+00		X 42	0.9152E-01	4.4029 **	0.1236	1.0006

TABLEAU 8 : LPOP3 = -0.34 + 0.85 Log(P1T + P2T) + 0.4 LLONG3

CARTE GRAPHIQUE

POP , LONG , POP / LONG , LOG , ET CARACTERISTIQUES DES BRANCHES

ABSCISSE = Y ESTIME ORDONNEE = RESIDU SYMBOLES . POINTS
+ POINTS SUPPLEMENTAIRES
* POINTS CONFONDUS



GRAPHIQUE 8 : LPOP3 = -0.34 + 0.85 Log(P1T+P2T) + 0.4 LLONG3

S. TOTALE	377.304
S. MODELE	250.771
S. RESIDUELLE	126.535

ECART-TYPE RESIDUEL	0.581
COEFFICIENT DE CORRELATION MULTIPLE	0.815
TEST GLOBAL D'AJUSTEMENT F(2, 375)	371.594 **

EQUATION DE LA REGRESSION

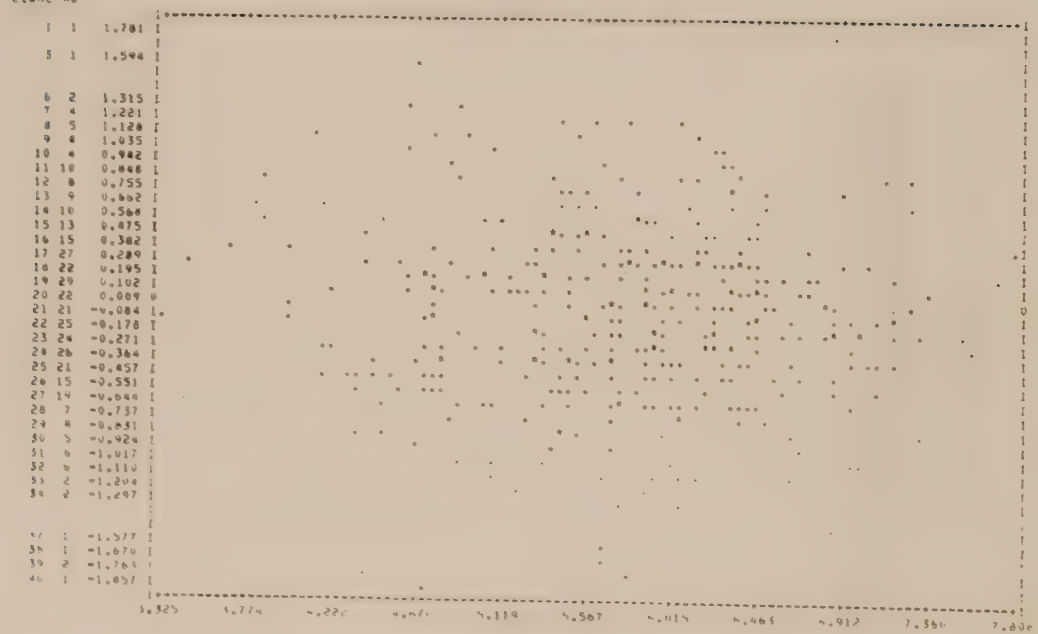
COEFFICIENT	NO. VARIABLE	ECART-TYPE	T. STUDENT	BETA	VIF
-0.1704E+00	CONST	0.3508E+00	-0.4857		
0.8251E+00	X 81	0.3074E-01	26.8384 **	0.8027	1.0003
0.5090E+00	X 42	0.9752E-01	5.2193 **	0.1561	1.0003

TABLEAU 10 : LPOP3 = -0.2 + 0.8 LP2T + 0.5 LLONG3

CARTE GRAPHIQUE

POP , LONG , POP / LONG , LON , ET CARACTERISTIQUES DES MNATICMES

ABSCISSE = Y ESTIME ORDONNEE = RESIDU SYMBOLES . POINTS
+ POINTS SUPPLEMENTAIRES
* POINTS CONFONDUS



GRAPHIQUE 10 : LPOP3 = -0.2 + 0.8 LP2T + 0.5 LLONG3

S. TOTALE	377.304
S. MODELE	199.360
S. RESIDUELLE	177.946

ECART-TYPE RESIDUEL	0.689
COEFFICIENT DE CORRELATION MULTIPLE	0.727
TEST GLOBAL D'AJUSTEMENT F(2, 375)	210.064 **

EQUATION DE LA REGRESSION

COEFFICIENT	NO. VARIABLE	ECART-TYPE	T. STUDENT	BETA	VIF
0.1683E+01	CONST	0.2821E+00	5.9680 **		
0.5289E+00	X 82	0.2700E-01	19.5895 **	0.6963	1.0044
0.5302E+00	X 85	0.7243E-01	7.3197 **	0.2602	1.0044

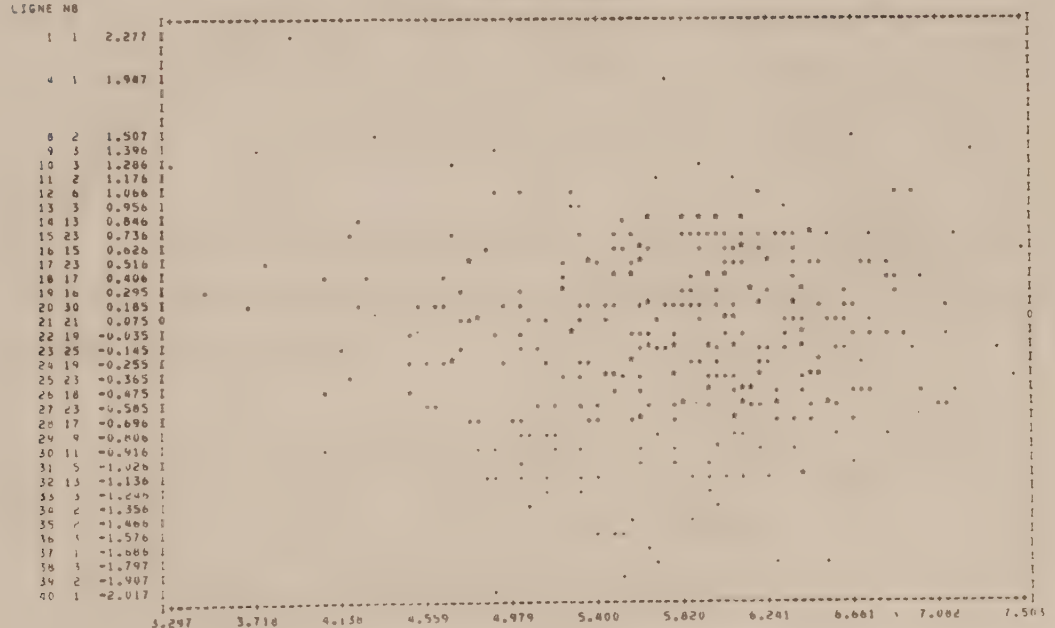
NORME DU VECTEUR DE REGRESSION : 0.7433
(DANS L'ESPACE DES PARAMETRES)

TABLEAU 12 : LPOP3 = 1.7 + 0.53 LP1T + LL2T

CARTE GRAPHIQUE

POP , LONG , POP / LONG , LOG , ET CARACTERISTIQUES DES BRANCHES

ABSCISSE = Y ESTIME ORDONNEE = RESIDU SYMBOLES . POINTS
+ POINTS SUPPLEMENTAIRES
= POINTS CONFONDUS



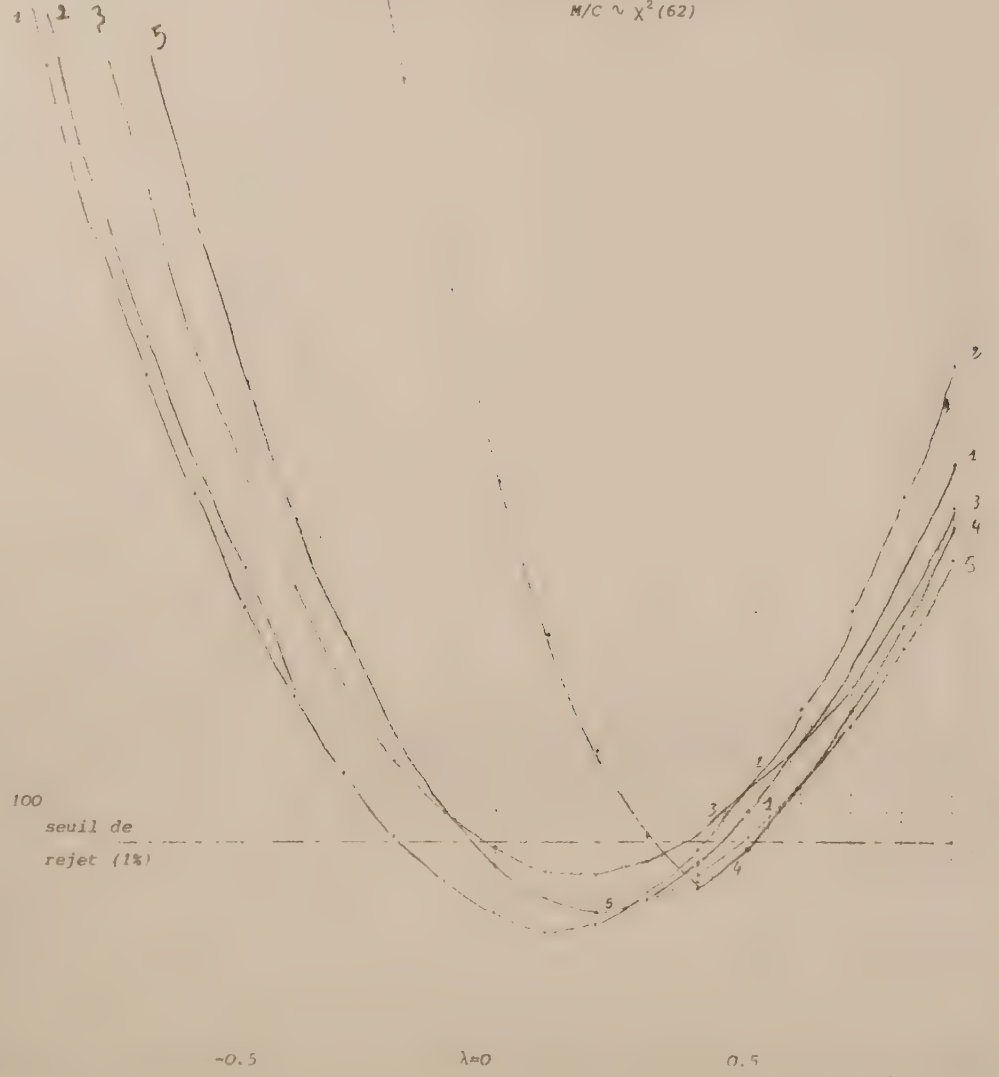
GRAPHIQUE 12 : LPOP3 = 1.7 + 0.53 LP1T + LL2T

GRAPHIQUE 14 : Variation de M en fonction de λ pour les populations
1 : PTOT, 2 : POP3, 3 : P2P, 4 : P1P et 5 : P12T

Modèle Rangée x travée

N=6 k=63 C=1.068

$M/C \sim \chi^2(62)$

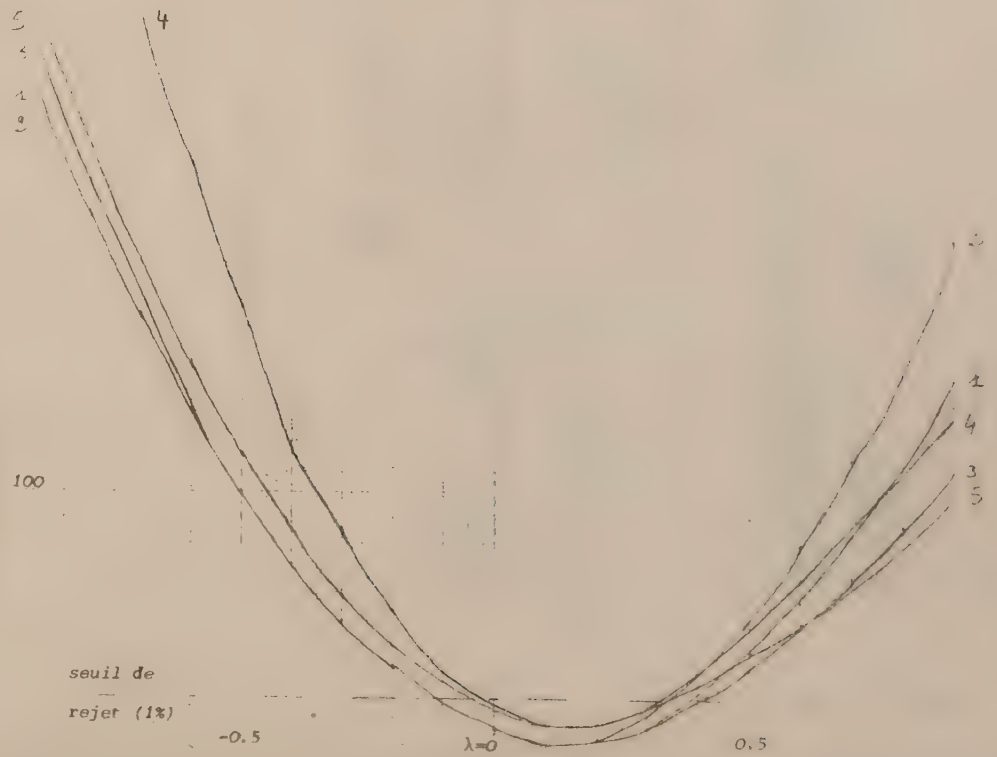


GRAPHIQUE 15 : Variation de M en fonction de λ pour les populations
1 : PTOT, 2 : POP3, 3 : P2P, 4 : P1P et 5 : P12T

Modèle Rangée

N=42 k=9 C=1.009

$M/C \sim \chi^2(8)$



A - II - 6) ANALYSE DE VARIANCE ET DE COVARIANCE

a) De nos connaissances a priori sur le caractère très contagieux de la diffusion de l'araignée rouge, nous pouvons nous attendre à ce que la position dans le verger ait une influence importante sur le degré d'infestation d'une branche.

Nous avons donc considéré les facteurs Rangée, Travée, Niveau et Face dans un modèle d'analyse de variance multidimensionnel, équilibré ([1], p.271).

Cette analyse suppose notamment que les erreurs sont normales, et que les variances intra strates sont égales (la strate étant définie par une combinaison des modalités de l'ensemble des facteurs étudiés).

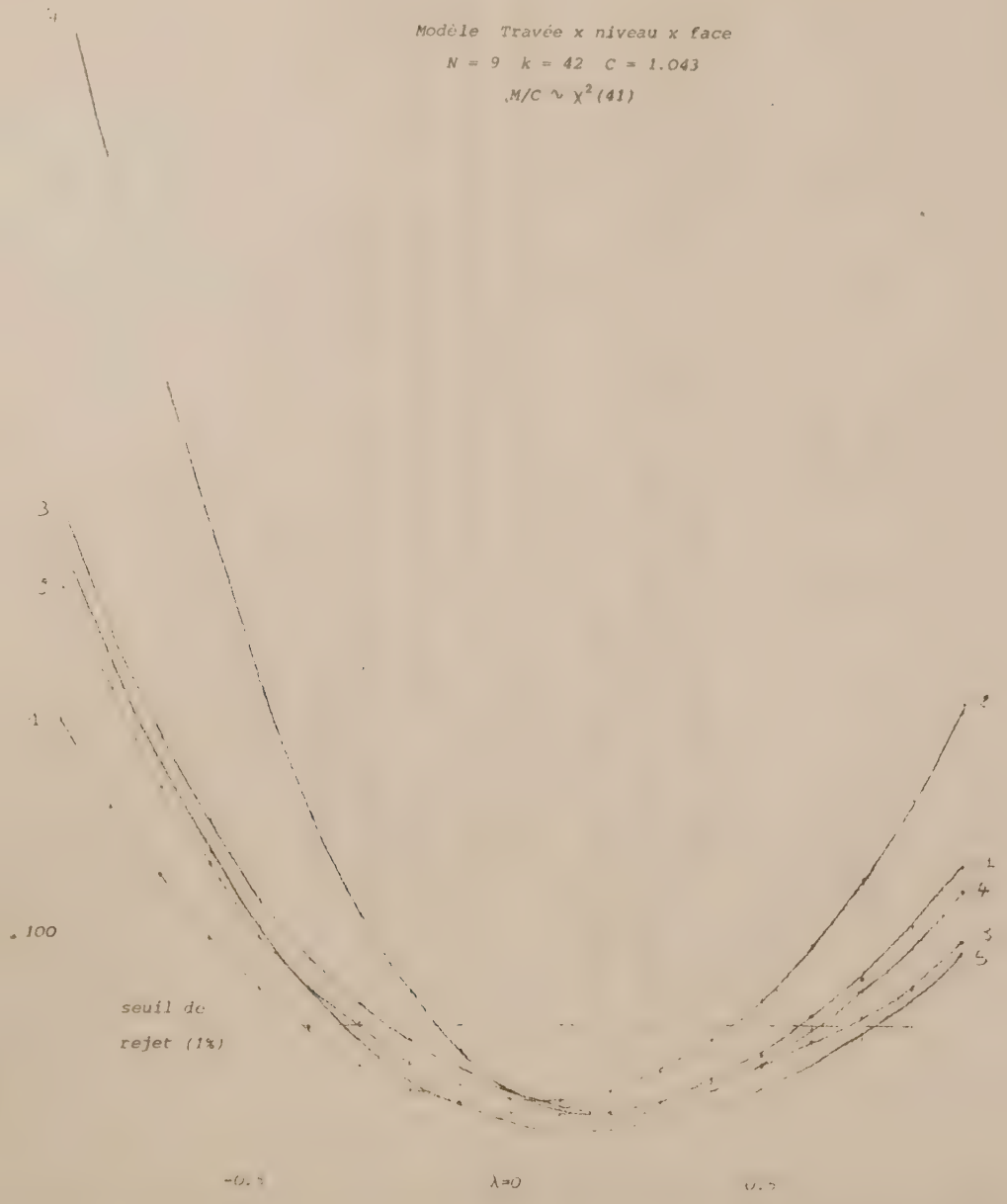
Ces hypothèses sont loin d'être vérifiées pour les différentes variables considérées. Nous avons donc considéré la série des transformations de Box et Cox ([2], p.522) pour les populations PTOT, POP3, P2P, P1P, P1M, P12M, P2M

$$X \rightarrow \begin{cases} \dfrac{(X+1)^{\lambda}}{\lambda} & si \quad \lambda \neq 0 \\ Log(x+1) & si \quad \lambda = 0 \end{cases}$$

Pour chaque transformation, nous avons appliqué l'analyse de variance à différents modèles (combinaison d'au plus 3 facteurs, le 4e servant comme répétition).

Notons k le nombre de strates (k = produit des niveaux maximums des facteurs inclus dans l'analyse), n le nombre d'individus dans chaque strate (constant, puisque le modèle est équilibré), et soit s_i^2, $i = 1...k$, la variance intra la strate i. Le test de BARTLETT ([2], p.132) repose sur la statistique M :

GRAPHIQUE 14 : Variation de M en fonction de λ pour les populations
1 : PTOT, 2 : POP3, 3 : P2P, 4 : P1P et 5 : P12T



GRAPHIQUE 14 : Variation de M en fonction de λ pour les populations
1 : PTOT, 2 : POP3, 3 : P2P, 4 : P1P et 5 : P12T

$$M = k(n-1)\ \mathrm{Log} \sum_{i=1}^{k} s_i^2/k - (n-1) \sum_{i=1}^{k} \mathrm{Log}\ s_i^2$$

$$\text{Posons } C = 1 + \frac{1}{3(k-1)} \left[\frac{k}{n-1} - \frac{1}{k(n-1)} \right]$$

Si les populations sont normales, et les variances s_i^2, i=1...k, égales, alors M/C suit, approximativement la loi $\chi^2(k-1)$ (l'approximation est très bonne même pour des échantillons à taille modérée). L'hypothèse d'homogénéité des variances est rejetée si $M/C > \chi^2_{\alpha,k-1}$. Si l'hypothèse est acceptée, la normalité de la distribution à l'intérieur de chaque groupe est en général vérifiée, la distribution de M étant très sensible à l'hypothèse de normalité ([5], p.63).

Pour la majorité des populations considérées (pas de données manquantes), et pour les différents modèles à 1 ou 3 facteurs, l'hypothèse d'homogénéité des variances est acceptée au seuil de 1% quand nous appliquons la transformation X → Log(X+1) - (cas λ = 0 dans G14).

Les populations de 1 an (P1P, P12T, P1T, P12T,...) se distinguent par l'allure de la statistique M qui varie plus vite avec λ.

b) Nous avons donc effectué l'analyse de variance ([1],p.301) de différents modèles sur les variables transformées par X → Log(X+1). Les résultats sont présentées sous 2 formes :

1) des représentations graphiques G15 ... G21 :

- *G15, G16 et G17 donnent les effets principaux des différents facteurs pour toutes les populations considérées (l'échelle est spécifique à chaque population).*

- *G18 ... G21, résument pour les populations PTOT, POP3, P2P, et P1P les résultats sur les modèles Rangées x Niveau x Face, Rangée, et Niveau x Face (échelle commune dans chaque graphe).*

Si m est la moyenne et σ^2 la variance intra estimées par le modèle, nous avons représenté une estimation de l'effet sur les variables non transformées : $\exp(m + \frac{1}{2}\ \sigma^2)$.

2) Deux sortes de tableaux récapitulatifs des principaux résultats :

- T13, T14 résument l'aspect univariable : pour chaque population, la décomposition de la variance et les tests correspondent à chaque composante du modèle (les variances intra-strates correspondent aux carrés moyens).

- T15, T16 donnent les matrices de covariances inter de chaque composante du modèle ainsi que la matrice intra et la matrice inter globale de tout le modèle.

G16 et T13 montrent que le facteur Travée n'est jamais significatif (sauf pour P1P au niveau de 5%). Par contre, la Rangée (G15) a un effet hautement significatif pour toutes les populations (remarquer la grande valeur du F correspondant, ainsi que le rapport des sommes des carrés associés, sur la somme des carrés totale est souvent voisin de 0.35).

Les effets du facteur Niveau sont variables suivant la Face (G17) de manière générale, (POP2 et POP3), nous constatons un gradient de bas en haut plus marqué sur la face nord que sur la face sud. Aussi la considération du facteur position dans l'arbre (Niveau x Face) parait plus convenable. Ce nouveau facteur est significatif mais bien moins que la Rangée (T14).

Cette première partie confirme donc l'aspect très contagieux de la distribution des oeufs d'hiver de l'araignée rouge. G18...G21 montre la présence d'un gradient dans l'effet de la Rangée : les Rangées 7, 8 et 9 sont très attaquées, alors que 1, 2 et 3 le sont très peu.

Les tableaux T15 et T16 montrent que les variations des effets moyens d'un facteur sont comparables d'une population à l'autre, sauf dans le cas du facteur Niveau (T15) et position dans l'arbre (Niveau x Face : T16), où P1P se distingue. (matrices de corrélation inter)

Les matrices de corrélation intra montrent que les variations des résidus, dans les 2 modèles, sont aussi très liées d'une population à l'autre, toujours en remarquant le rôle particulier de LP1P.

Pour élucider ce fait, nous avons donc considéré de nouveau les caractéristiques architecturales de la branche comme variables explicatives, en tenant compte des résultats de l'analyse de variance.

Avant de passer aux résultats de cette étude, notons que le rôle particulier joué par la population sur les rameaux de 1 an pourrait n'être dû qu'aux effets de la taille sur l'arbre. En effet, la taille est en général particulièrement sévère sur les rameaux du niveau 3 et dans l'axe du tronc pour empêcher l'arbre de croitre en hauteur. Mais aussi c'est à ce niveau qu'on pourrait trouver les bois les plus vigoureux comme réaction de l'arbre.

Ceci nous amène donc à nous poser le problème du lien entre les longueurs et la position dans le verger. Les résultats de l'analyse de variance sur le modèle Rangée x (niveau x face), pour les longueurs LLTOT, LLONG3, LL2P, LL1P sont résumées dans G22.

Les effets des 2 facteurs Rangée et Niveaux Face sont significatifs au niveau de 1% pour les différentes longueurs sauf LONG3.

Pour les effets du facteur Niveau x Face, les résultats sont compréhensibles biologiquement pour L1P (les longueurs sont plus grandes sur la face sud que sur la face nord, et il y a un gradient positif de bas en haut). Pour L2P, les résultats peuvent s'interpréter par l'effet de la taille qui serait plus sévère au niveau moyen et rendrait les deux faces symétriques. Ce facteur n'est pas significatif pour LONG3 et les effets sur LTOT rendent compte de la complexité de l'ensemble.

Pour les Rangées, le même gradient est dans l'ensemble présent sur toutes les longueurs et concorde partiellement à celui constaté sur les populations, sauf pour les Rangées 3, 6 et 8.

Remarquons finalement que les moyennes générales (de 3 à 5) sont très importantes comparées aux effets moyens ajustés (de -0.3 à 0.3). Les valeurs du R^2 sont néanmoins assez importantes (0.2 pour L1P, 0.32 pour L2P et LTOT).

ANALYSES DE VARIANCE

CARACTERE NO : 27 (1) LPTOT

SOURCE DE VARIATION		SOMME DES CARRES	DLIB	CARRE MOYEN	F	
FACTEUR	RANGEE	0.13445906E+03	8	0.16807373E+02	35.490	**
FACTEUR	TRAVEE	0.53671856E+01	6	0.89453089E+00	1.889	
FACTEUR	NIVEAU	0.21240143E+02	2	0.10620071E+02	22.425	**
FACTEUR	FACE	0.21765721E+00	1	0.21765721E+00	0.460	
ERREUR		0.17048901E+03	360	0.47358054E+00		
TOTAL		0.33177295E+03	377			

CARACTERE NO : 28 (2) LPOP3

SOURCE DE VARIATION		SOMME DES CARRES	DLIB	CARRE MOYEN	F	
FACTEUR	RANGEE	0.12907120E+03	8	0.16133896E+02	28.267	**
FACTEUR	TRAVEE	0.50556908E+01	6	0.84261513E+00	1.476	
FACTEUR	NIVEAU	0.37217606E+02	2	0.18608795E+02	32.603	**
FACTEUR	FACE	0.90993285E-01	1	0.90993285E-01	0.159	
ERREUR		0.20547948E+03	360	0.57077628E+00		
TOTAL		0.37691479E+03	377			

CARACTERE NO : 29 (3) LP2P

SOURCE DE VARIATION		SOMME DES CARRES	DLIB	CARRE MOYEN	F	
FACTEUR	RANGEE	0.12472334E+03	8	0.15590418E+02	27.002	**
FACTEUR	TRAVEE	0.57079916E+01	6	0.95133191E+00	1.648	
FACTEUR	NIVEAU	0.17982788E+02	2	0.89913940E+01	15.573	**
FACTEUR	FACE	0.17746389E+00	1	0.17746389E+00	0.307	
ERREUR		0.20785303E+03	360	0.57736951E+00		
TOTAL		0.35644458E+03	377			

CARACTERE NO : 31 (4) LP1P

SOURCE DE VARIATION		SOMME DES CARRES	DLIB	CARRE MOYEN	F	
FACTEUR	RANGEE	0.18888420E+03	8	0.23610519E+02	28.274	**
FACTEUR	TRAVEE	0.12011065E+02	6	0.20018435E+01	1.719	
FACTEUR	NIVEAU	0.20673462E+02	2	0.10336731E+02	8.875	**
FACTEUR	FACE	0.12319651E+01	1	0.12319651E+01	1.058	
ERREUR		0.41923828E+03	360	0.11645508E+01		
TOTAL		0.64203882E+03	377			

TABLEAU 13 : Aspects unidimensionnels des résultats de l'analyse de variance sur le modèle (1) :

$$P_{ijk\ell} = m + R_i + T_j + N_k + F_\ell + \varepsilon_{ijk\ell}$$

avec i, j, k et ℓ désignant respectivement une modalité de la rangée, la travée, le niveau et la face, et P une des populations LPTOT, LPOP3, LP2P, LP1P (transformées)

ANALYSES DE VARIANCE

CARACTERE NO : 27 (1) LPTOT

SOURCE DE VARIATION			SOMME DES CARRES	DLIB	CARRE MOYEN	F	
FACTEUR	RANGEE		0.13445891E+03	8	0.16807358E+02	39.197	**
FACTEUR	NIV*F		0.24606857E+02	5	0.49213715E+01	11.477	**
INTERACTION	RANGEE-	NIV*F	0.33161133E+02	40	0.82902831E+00	1.933	**
ERREUR			0.13892865E+03	324	0.42879212E+00		
TOTAL			0.33115552E+03	377			

CARACTERE NO : 28 (2) LPOP3

SOURCE DE VARIATION			SOMME DES CARRES	DLIB	CARRE MOYEN	F	
FACTEUR	RANGEE		0.12907103E+03	8	0.16133865E+02	31.077	**
FACTEUR	NIV*F		0.40853607E+02	5	0.81707211E+01	15.738	**
INTERACTION	RANGEE-	NIV*F	0.39685608E+02	40	0.99214017E+00	1.911	**
ERREUR			0.16820631E+03	324	0.51915526E+00		
TOTAL			0.37781641E+03	377			

CARACTERE NO : 29 (3) LP2P

SOURCE DE VARIATION			SOMME DES CARRES	DLIB	CARRE MOYEN	F	
FACTEUR	RANGEE		0.12472318E+03	8	0.15590397E+02	27.224	**
FACTEUR	NIV*F		0.19825150E+02	5	0.39650297E+01	6.924	**
INTERACTION	RANGEE-	NIV*F	0.24054596E+02	40	0.60136485E+00	1.050	
ERREUR			0.18554823E+03	324	0.57267970E+00		
TOTAL			0.35415112E+03	377			

CARACTERE NO : 31 (4) LP1P

SOURCE DE VARIATION			SOMME DES CARRES	DLIB	CARRE MOYEN	F	
FACTEUR	RANGEE		0.18888394E+03	8	0.23610489E+02	21.525	**
FACTEUR	NIV*F		0.26783401E+02	5	0.53566799E+01	4.883	**
INTERACTION	RANGEE-	NIV*F	0.68730881E+02	40	0.17182713E+01	1.566	*
ERREUR			0.35539697E+03	324	0.10969038E+01		
TOTAL			0.63979517E+03	377			

TABLEAU 14 : Aspects unidimensionnels des résultats de l'analyse de variance sur le modèle (2) :

$$P_{ijk\ell} = m + R_i + (NF)_{k\ell} + (RNF)_{ik\ell} + \varepsilon_{ijk\ell}$$

avec i, j, k et ℓ désignant respectivement une modalité de la rangée, la travée, le niveau et la face, et P une des populations LPTOT, LPOP3, LP2P, LP1P (transformées)

MATRICES INTER DU FACTEUR RANGEE

Corrélations

	27	28	29	31
27	1000			
28	997	1000		
29	993	989	1000	
31	990	979	985	1000

MATRICES INTER DU FACTEUR TRAVEE

Corrélations

	27	28	29	31
27	1000			
28	907	1000		
29	835	588	1000	
31	651	317	794	1000

MATRICES INTER DU FACTEUR NIVEAU

Corrélations

	27	28	29	31
27	1000			
28	998	1000		
29	998	993	1000	
31	-433	-483	-379	1000

MATRICES INTER DU FACTEUR FACE

Corrélations

	27	28	29	31
27	1000			
28	1000	1000		
29	1000	1000	1000	
31	1000	1000	1000	1000

MATRICES INTRA (W)

Corrélations

	27	28	29	31
27	1000			
28	936	1000		
29	780	607	1000	
31	610	463	544	1000

MATRICES INTER TOTALES (B)

Corrélations

	27	28	29	31
27	1000			
28	988	1000		
29	988	968	1000	
31	814	728	830	1000

TABLEAU 15 : suite du tableau 13 : Aspects multivariables

MATRICES INTER DU FACTEUR RANGEE

Corrélations

	27	28	29	31
27	1000			
28	997	1000		
29	993	989	1000	
31	990	979	985	1000

MATRICES INTER DU FACTEUR NIV::F

Corrélations

	27	28	29	31
27	1000			
28	993	1000		
29	985	975	1000	
31	-182	-269	-196	1000

MATRICES INTER DE L'INTERACTION : RANGEE - NIV::F

Corrélations

	27	28	29	31
27	1000			
28	960	1000		
29	823	730	1000	
31	677	545	579	1000

MATRICES INTRA (W)

Corrélations

	27	28	29	31
27	1000			
28	930	1000		
29	778	589	1000	
31	591	435	539	1000

MATRICES INTER TOTALES (B)

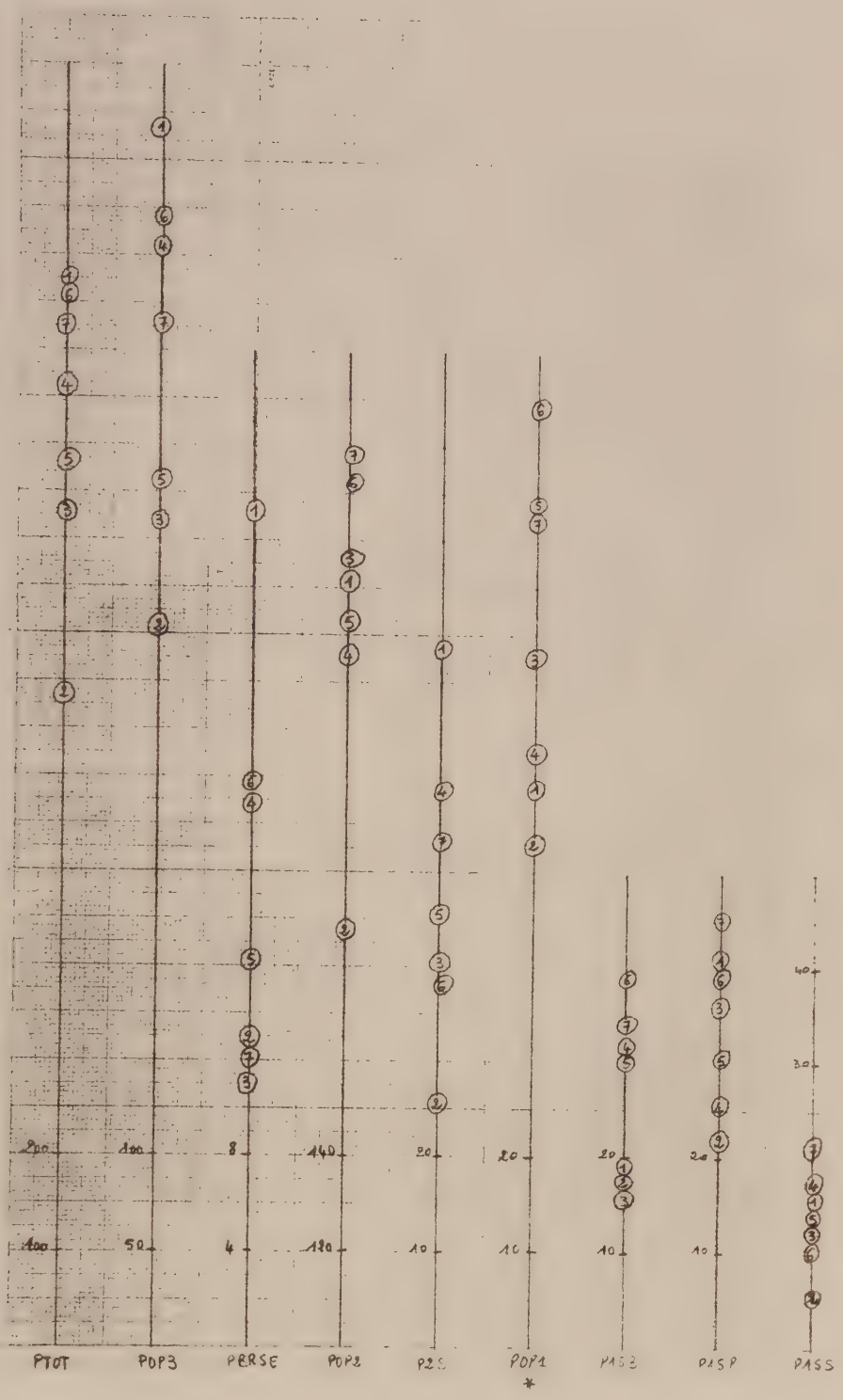
Corrélations

	27	28	29	31
27	1000			
28	985	1000		
29	964	935	1000	
31	793	706	778	1000

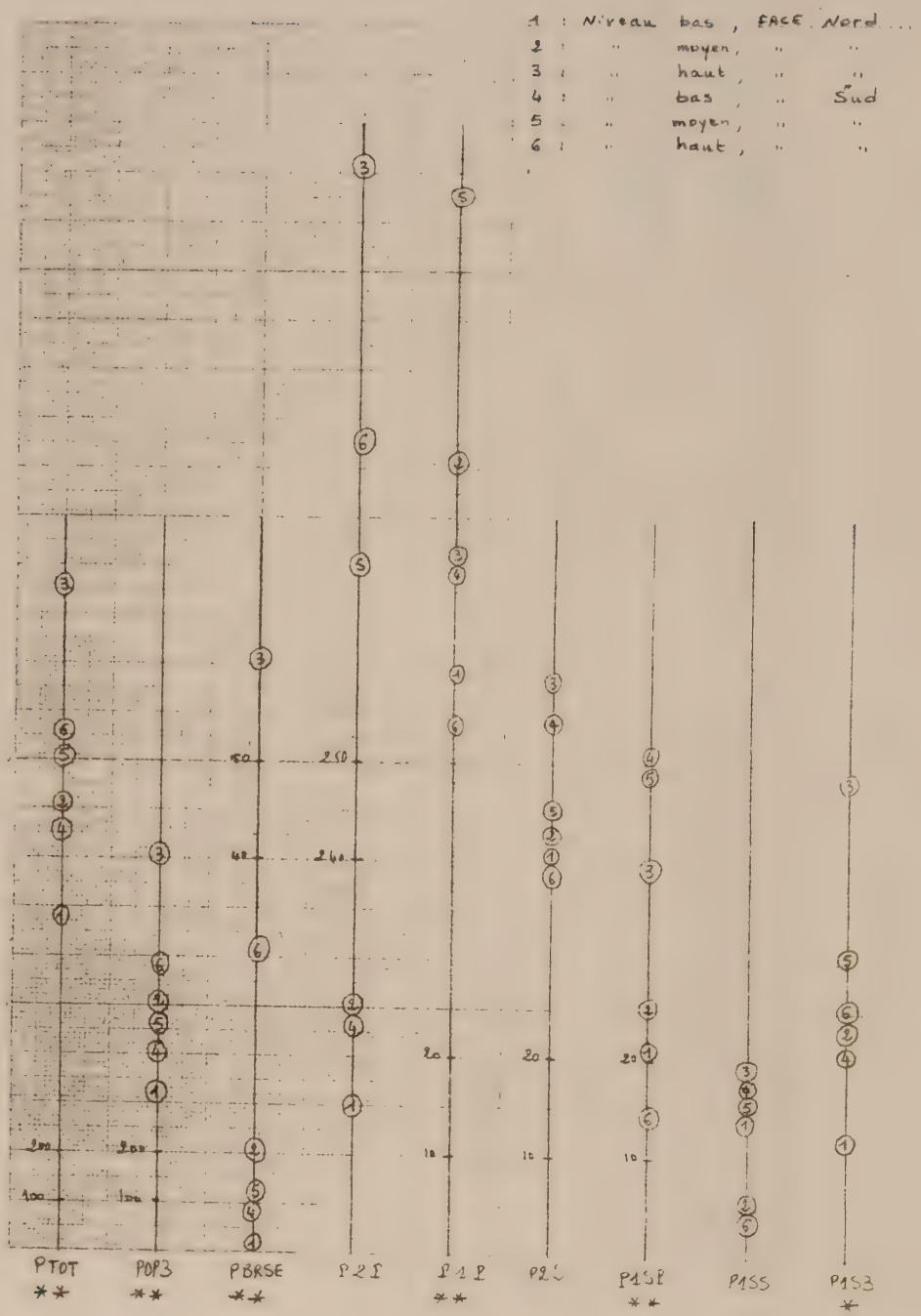
TABLEAU 16 : suite de T14 : Aspects multivariates

GRAPHIQUE 15 : FACTEUR RANGEE

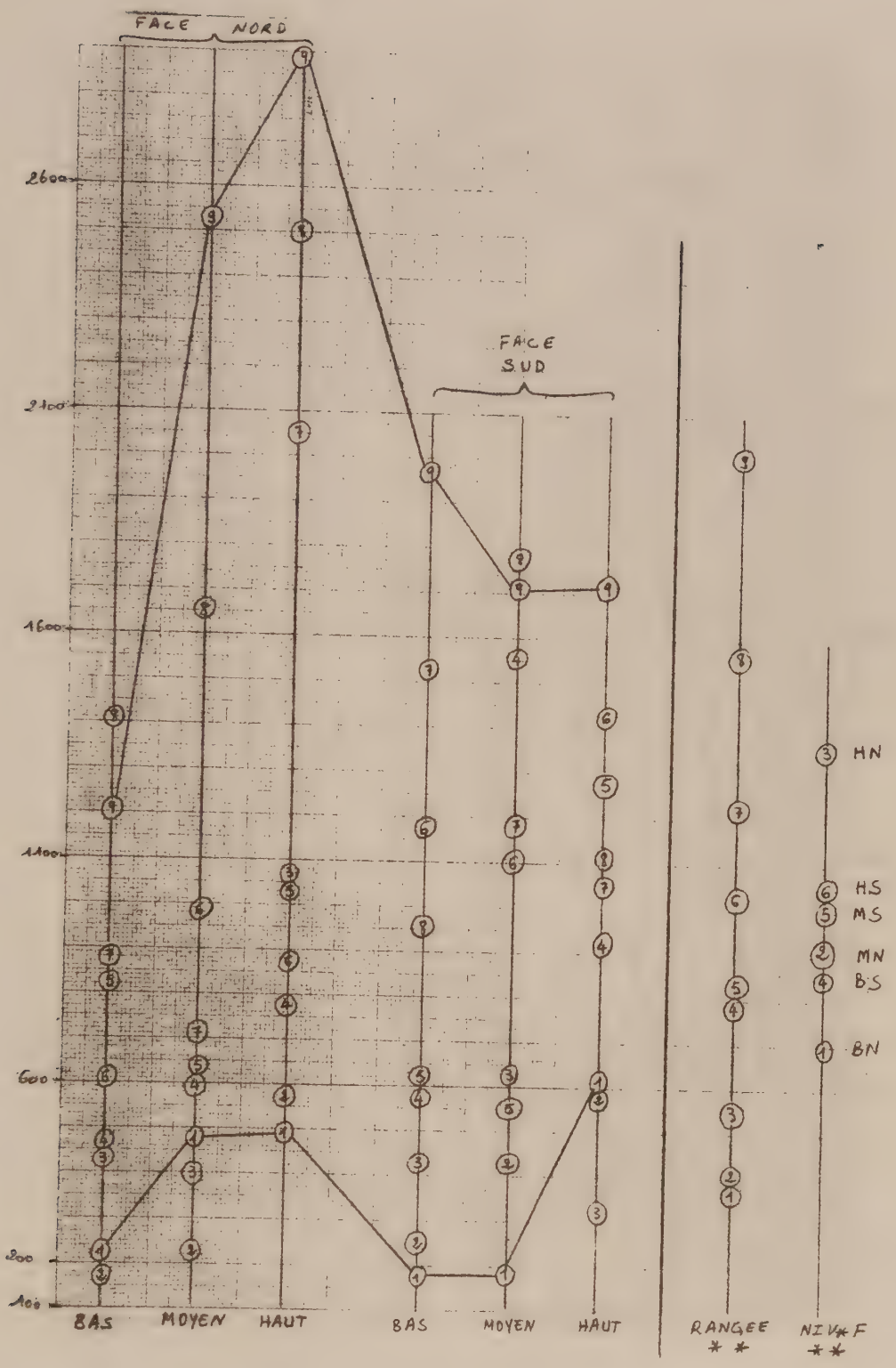
GRAPHIQUE 16 : FACTEUR : TRAVEE



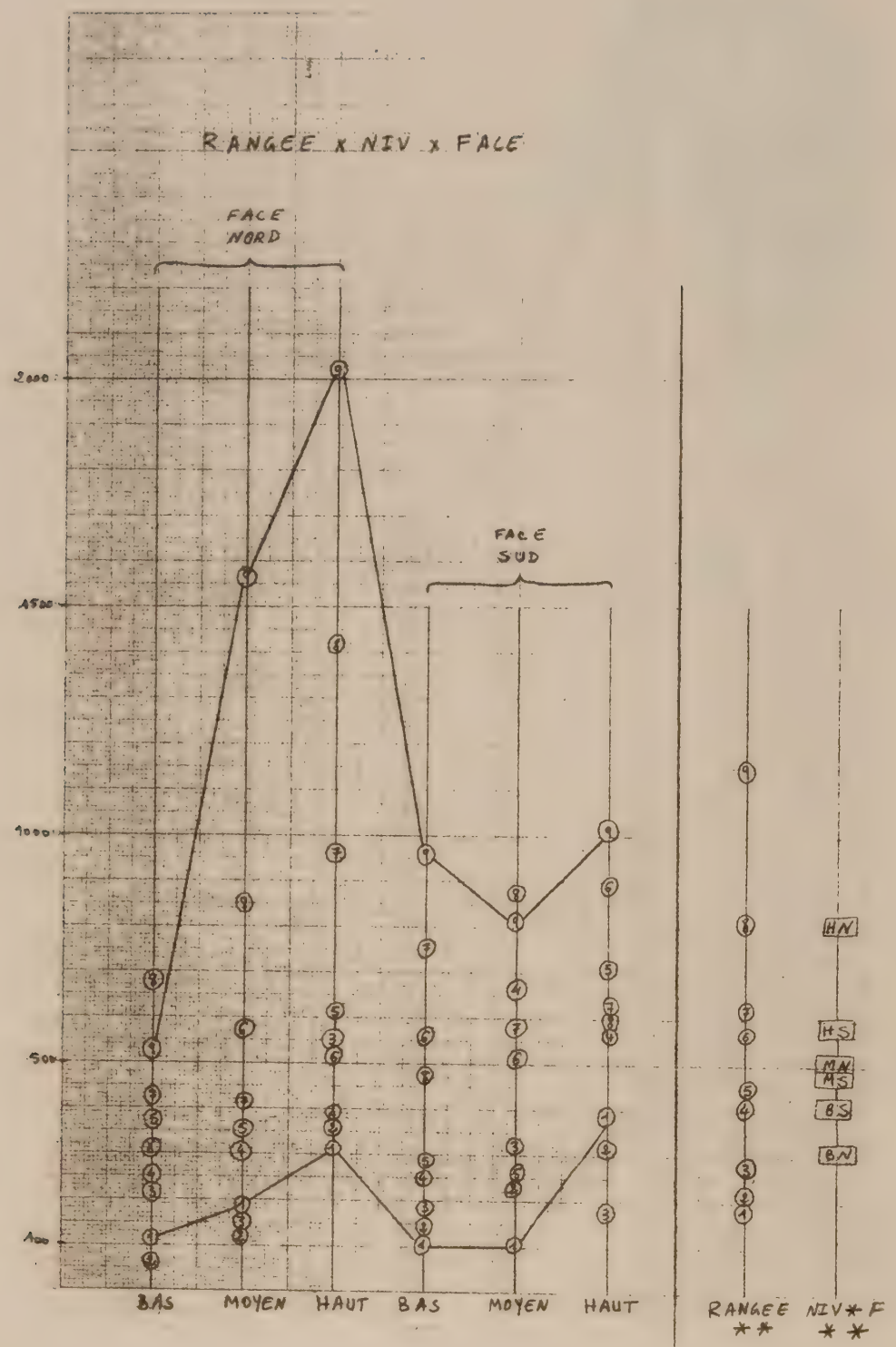
GRAPHIQUE 17 : FACTEUR : NIVEAU × FACE



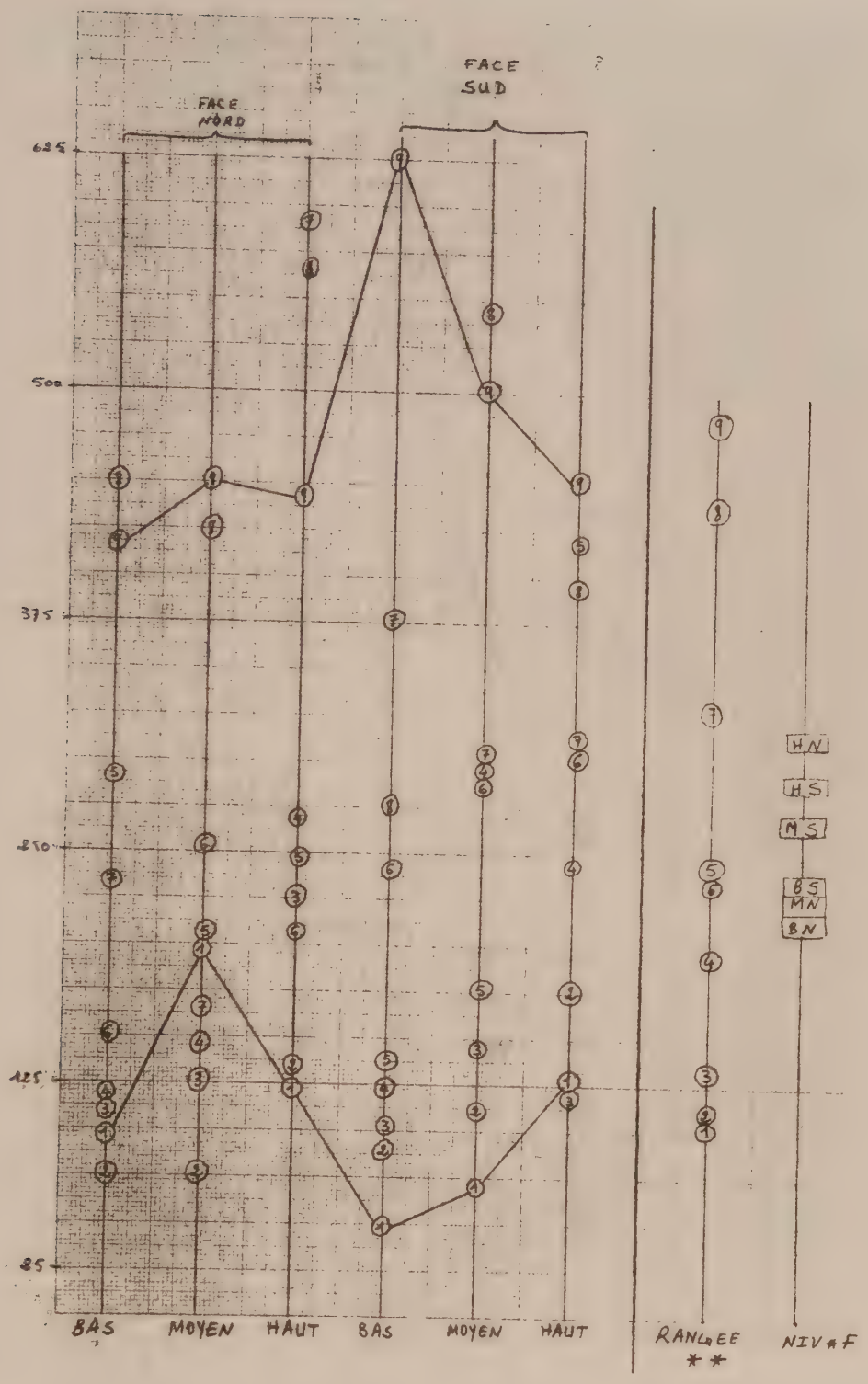
GRAPHIQUE 18 : PTOT



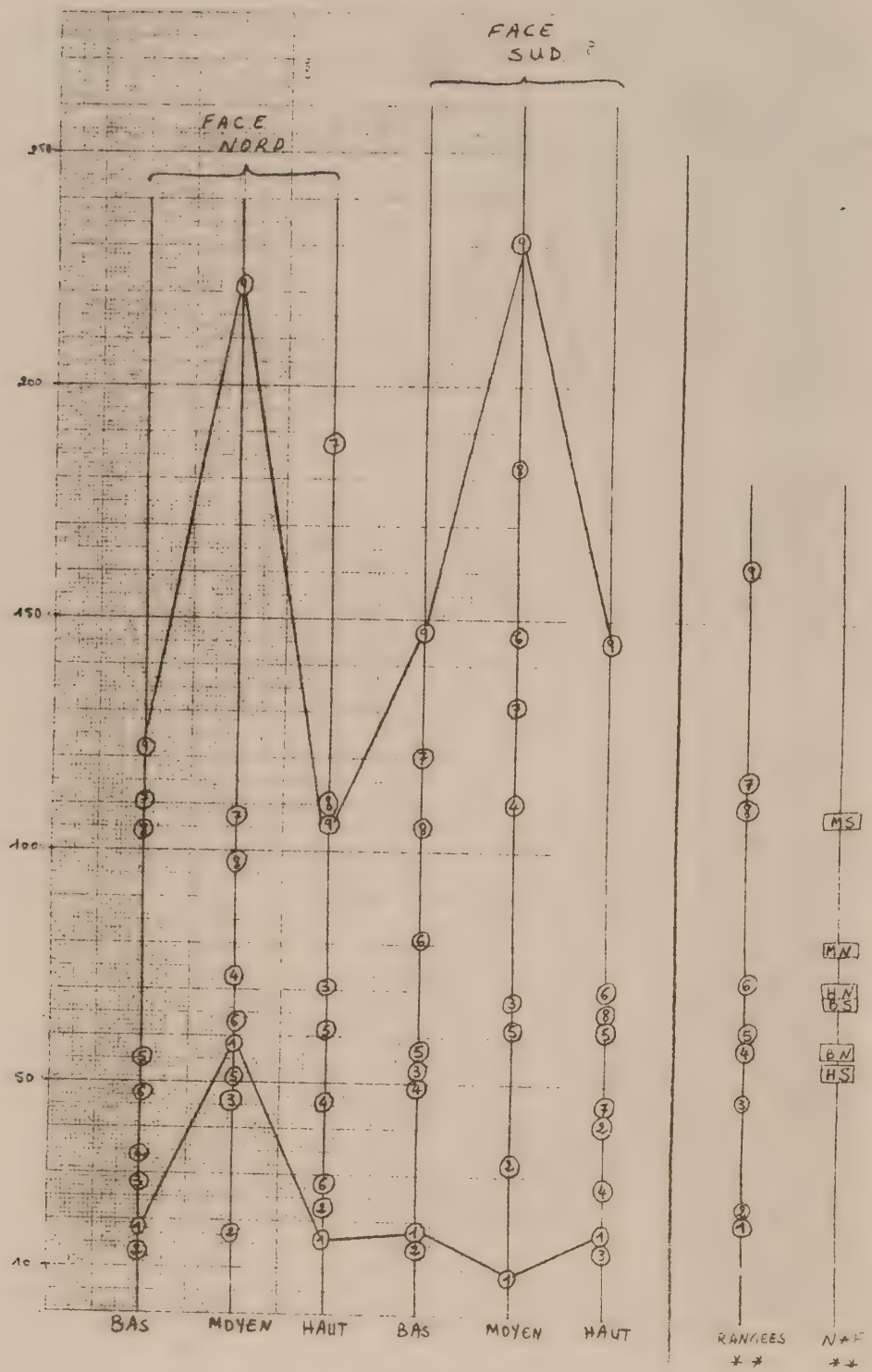
GRAPHIQUE 19 : POP3



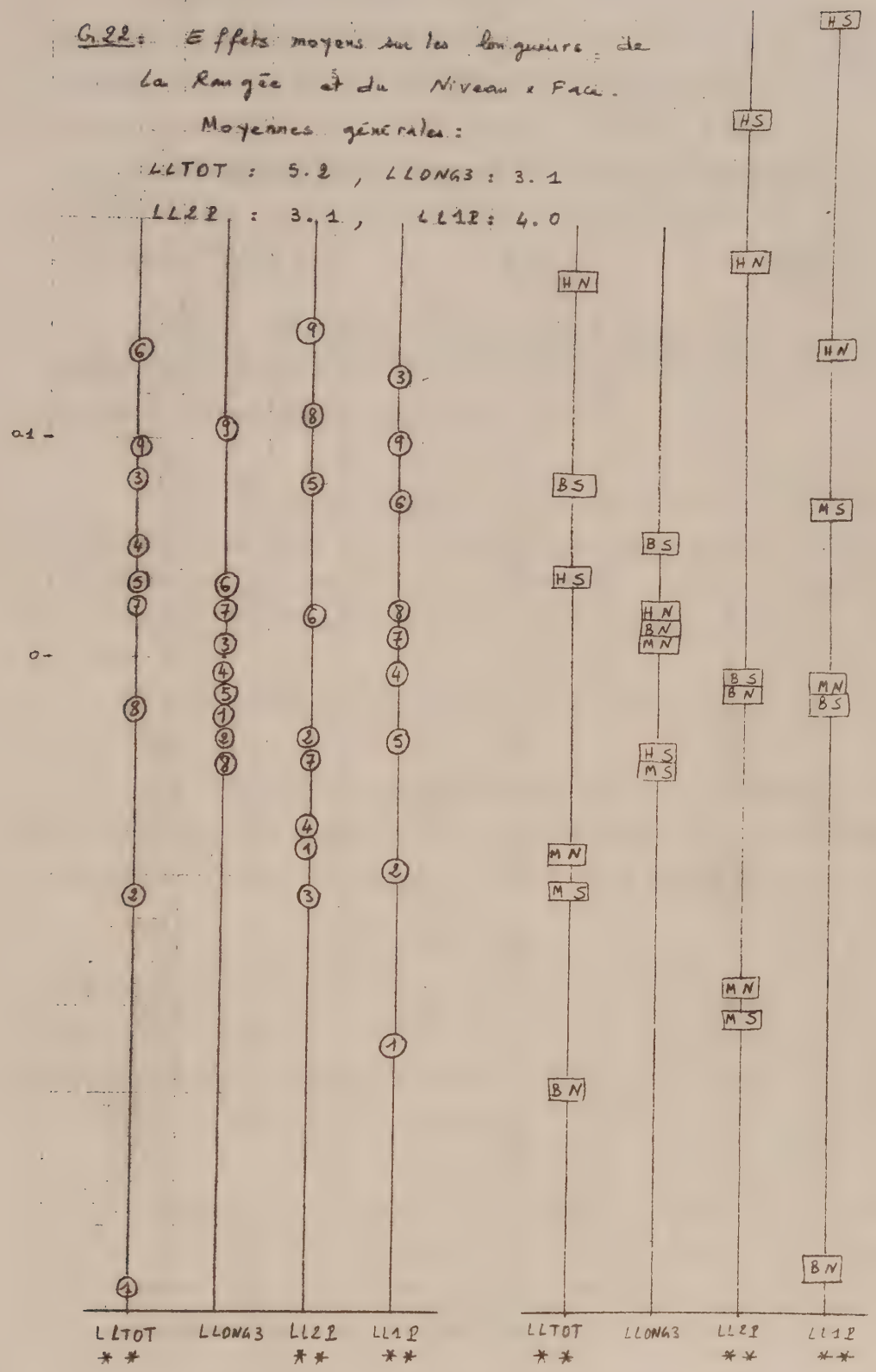
GRAPHIQUE 20 : P2P



GRAPHIQUE 21 : P1P



GRAPHIQUE 22



GRAPHIQUE 23

G23 :

$$L2D23 = 1.65 + 0.80\ LLTOT + (R) + (NF) + (R \times NF)$$

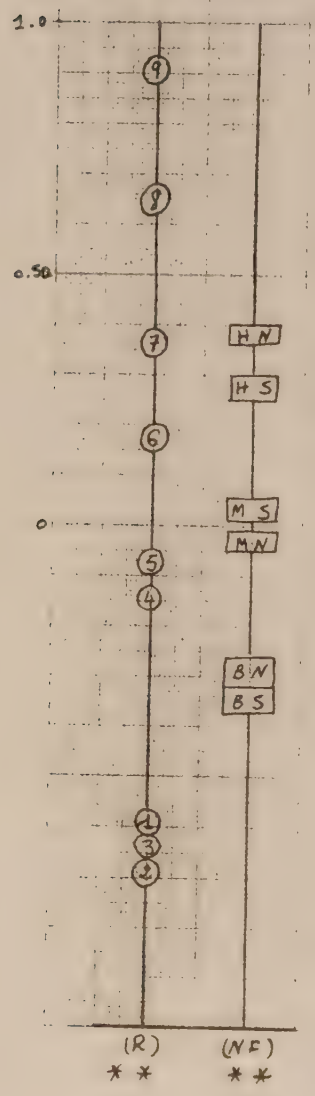
$$L221 = 1.51 + 0.77\ LL23T + (R) + (NF) + (R \times NF)$$

$$L211 = -0.8 + 1.11\ LL1I + (R) + (NF) + (R \times NF)$$

L2D23

$R^2 = 0.63$

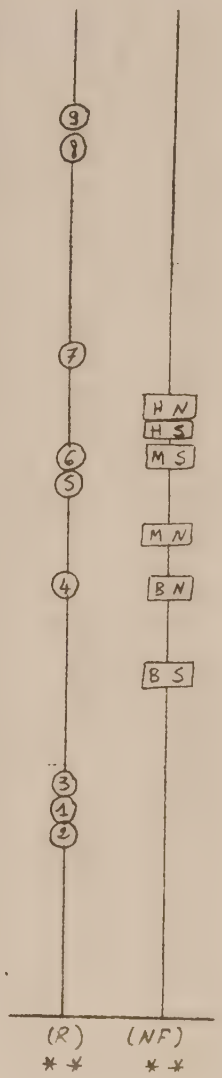
$F(54, 323) = 10.18$



L221

$R^2 = 0.53$

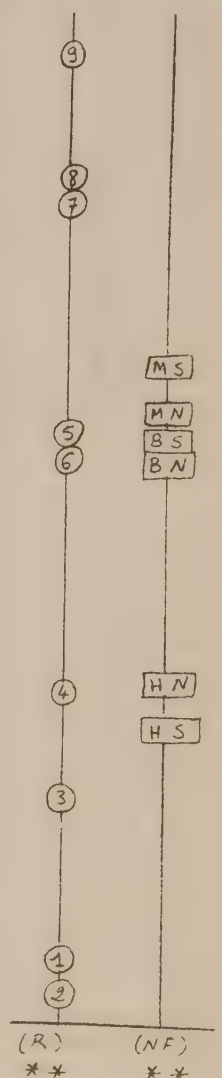
$F(54,323) = 6.75$



L211

$R^2 = 0.51$

$F(54,323) = 5.69$



c) Analyse de covariance

Nous avons d'abord introduit les caractéristiques architecturales des branches comme covariables, dans le modèle d'analyse de variance Rangée x Niveau x face. ([1], p.459). Les résultats sont résumés dans G23.

Les hypothèses d'absence des effets de chacune des composantes du modèle, ainsi que celle de nullité du coefficient de régression sont rejetées au niveau 1%.

Le même gradient sur les effets de la Rangée est présent pour les 3 populations considérées et concorde dans l'ensemble avec celui constaté lors de l'analyse de variance (voir G15).

Pour le niveau x face, la similitude entre les gradients dans le cas de l'analyse avec ou sans covariables (POP3 et P2P) subsiste. La covariable met plus en relief les effets du niveau (gradient positif de bas en haut), et estompe l'effet de la Face (comparer avec G17).

Les résultats concernant P1P montrent un phénomène particulier :

- l'analyse de variance (G17) montre une inversion du gradient entre bas-moyen et moyen-haut. La même analyse sur les longueurs (G22) met en évidence un gradient positif (assez fort) bas-moyen-haut.

- les résultats de l'analyse de covariance (G23) rend compte d'un effet important de la longueur (coefficient = 1.11), et d'un effet négatif marqué de la position haute sur le bois de 1 an (notons qu'il s'agit des parties les moins protégées contre le vent et la pluie, et que les comptages sont effectués à la fin de l'hiver).

G24 et G25 présentent les résultats quand nous considérons comme covariables seulement les populations sur les rameaux plus jeunes (G24) ou en même temps les populations sur les rameaux plus jeunes et les longueurs (G25). Les mêmes observations que dans G23 sont valables.

Dans G24, nous constatons cependant qu'il y a surtout une opposition entre le niveau haut et les autres. L'apport des longueurs dans G25 est très

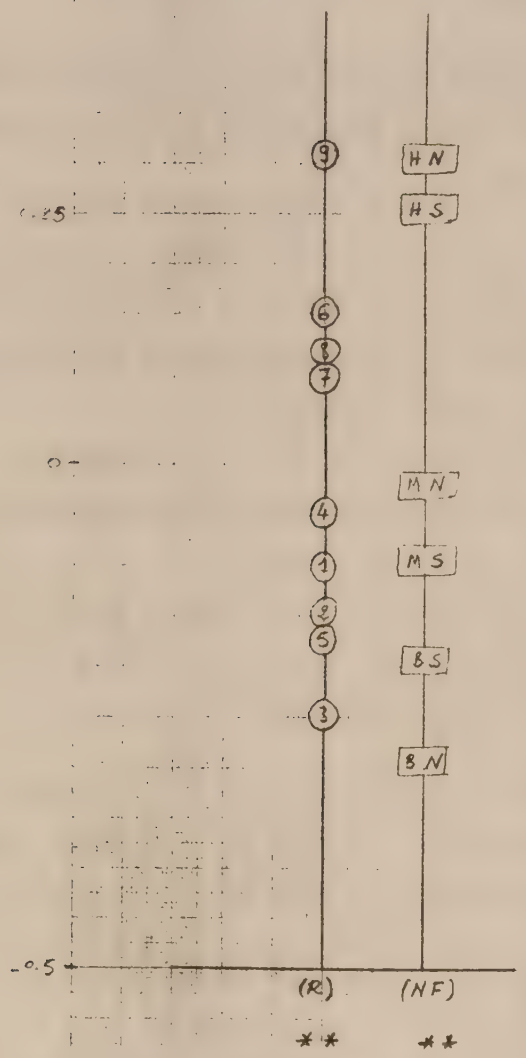
GRAPHIQUE 24

G24 : LPOP3 = 1.7 + 0.71 Log(P1T + P2T) + (R) + (NF) + (R * NF)

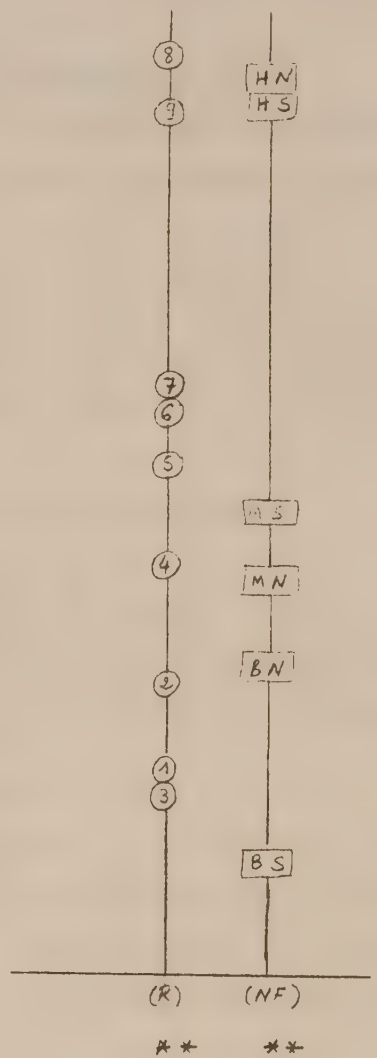
LP2P = 3.3% + 0.44 LP12PT + (R) + (NF) + (R * NF)

LPOP3 : R^2 = 0.78 F(54,323) = 21.2

LP2P : R^2 = 0.67 F(54,323) = 12.14



LPOP3



LP2P

négligeable même si le test de nullité du coefficient de régression correspondant est significatif à 1%.

Les différents résultats sont résumés dans T18, où sont rapportées le rapport des sommes des carrés du modèle sur la somme des carrés totale (R^2), ainsi que la valeur de la statistique servant à effectuer le test global d'ajustement, qui sont tous significatifs au niveau de 1%.

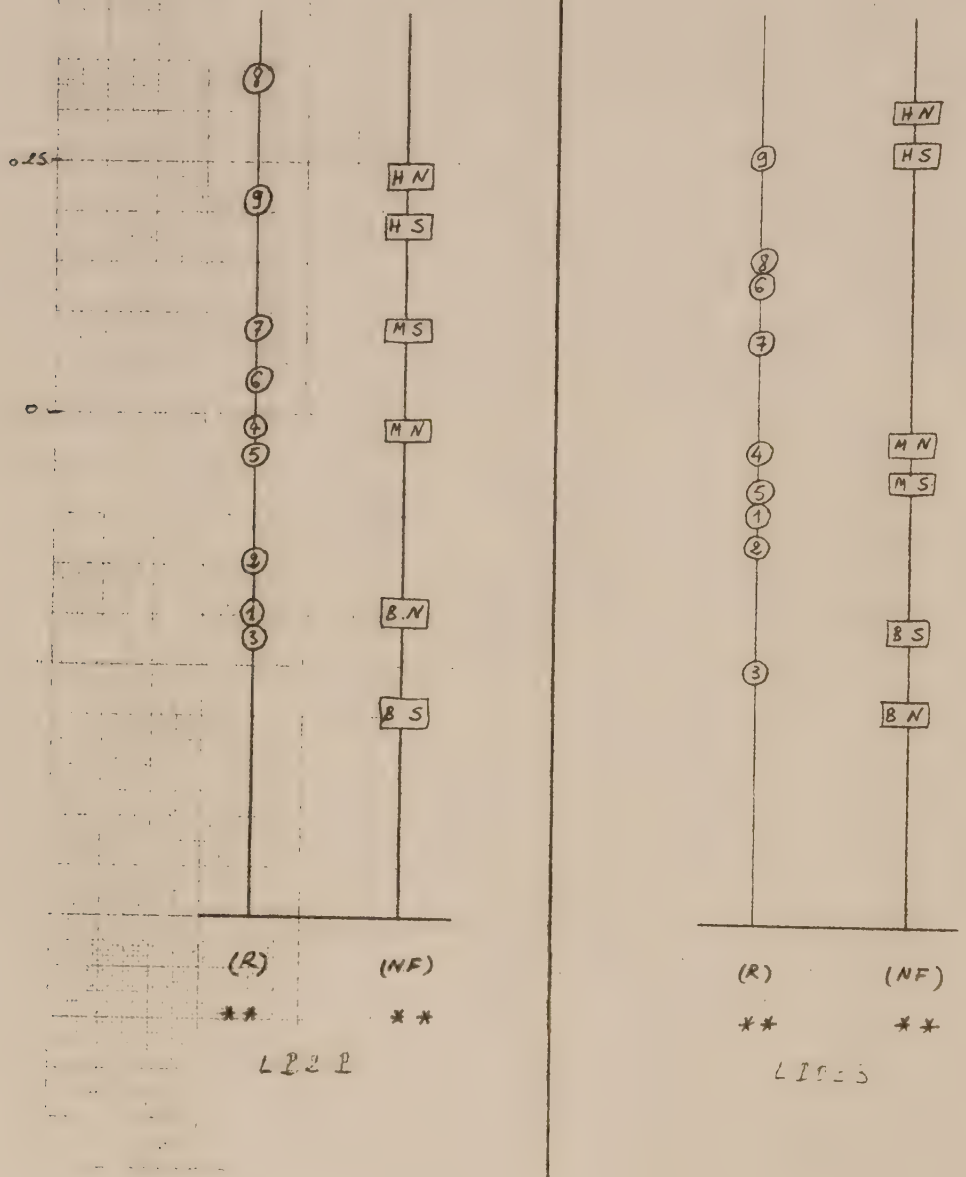
GRAPHIQUE 25

G25 : L2Q23 = 0.31 + (R) + (NF) + (RxNF) + 0.72 log(21T + 2ET) + 0.44 LLQH43

L222 = 0.08 + (R) + (NF) + (RxNF) + 0.47 L212PT + 0.76 LL21

L222 : R^2 = 0.73 ; F(55,332) = 15.8

L2Q23 : R^2 = 0.79 ; F(55,332) = 22



	RELATION	R^2	Test global d'ajustement	
(1)	LPOP3 = -0.38 + 1.2 LLTOT	0.22	F(1,376) = 109	**
(2)	= 0.91 + 0.86 Log(P1T+P2T)	0.68	F(1,376) = 835	**
(3)	= -0.34 + 0.85 Log(P1T+P2T) + 0.4 LLONG3	0.70	F(2,375) = 448	**
(4)	= m + (R) + (NF) + (R * NF)	0.56	F(53,324) = 7.8	**
(5)	= m + (R) + (NF) + (R * NF) + 0.80 LLTOT	0.63	F(54,323) = 10.2	**
(6)	= m + (R) + (NF) + (R * NF) + 0.71 Log(P1T+P2T)	0.78	F(54,323) = 21.2	**
(7)	= m + (R) + (NF) + (R * NF) + 0.72 Log(P1T+P2T) + 0.44 LLONG3	0.79	F(55,322) = 22	**
(8)	LP2P = 0.1 + 1.06 LL2PT	0.18	F(1,376) = 85	**
(9)	= 3.2 + 0.49 LP12PT	0.47	F(1,376) = 346	**
(10)	= -0.07 + 0.52 LP12PT + 0.97 LL2PT	0.64	F(2,375) = 332	**
(11)	= m + (R) + (NF) + (R * NF)	0.47	F(53,324) = 4.6	**
(12)	= m + (R) + (NF) + (R * NF) + 0.77 LL2PT	0.53	F(54,323) = 6.8	**
(13)	= m + (R) + (NF) + (R * NF) + 0.44 LP12PT	0.67	F(54,323) = 12.1	**
(14)	= m + (R) + (NF) + (R * NF) + 0.47 LP12PT + 0.76 LL2P	0.73	F(55,322) = 15.8	**
(15)	LP1P = - 1.5 + 1.3 LL1P	0.14	F(1,376) = 63	**
(16)	= m + (R) + (NF) + (R * NF)	0.44	F(53,324) = 4.8	**
(17)	= m + (R) + (NF) + (R * NF) + 1.11 LL1P	0.51	F(54,323) = 5.7	**

TABLEAU 18 : Résumé des différentes relations

A-II-7) CONCLUSIONS

a) Dans les différentes analyses, nous avons été amenés à considérer la transformation $X \rightarrow Log(X+1)$ sur toutes les populations :

- le tableau T3 montre que le comportement des variables transformées est proche de la normale.

- les graphes des résidus en fonction des estimées des différentes régressions montre que cette transformation est adéquate pour que les résidus puissent être considérés comme des réalisations indépendantes d'une variable aléatoire normale, centrée et de variance constante.

- dans les différents modèles d'analyse de variance, les graphes G14 montrent que cette transformation est convenable dans la majorité des cas.

Nous supposerons donc que chaque population suit une loi lognormale ou une combinaison de lois lognormales.

b) Les meilleures relations explicatives de la population sur un rameau sont obtenues quand nous tenons compte de la position dans le verger, des populations sur les rameaux plus jeunes et des caractéristiques architecturales de la branche, notamment la longueur du rameau considéré.

Dans tous les cas, l'apport des longueurs n'améliore pas de beaucoup le modèle sans longueurs, même si les coefficients de la régression sont significativement non nuls au niveau de 1%.

c) Les figures 7, 8 et 9 donnent une idée sur la grande variabilité des relations entre les différentes populations. D'autre part, le pourcentage moyen (Fig.8) des oeufs déposés sur les rameaux de 3 ans et de 2 ans principal sont très élevés (53% et 26% respectivement) ceci est vrai même si nous ne considérons que les 56 branches telles que L1SS et L1S3 non nuls.

Les figures 1...5 donnent une idée sur les relations entre populations pour des bois de 1 à 12 ans. Le nombre d'oeufs déposés sur un vieil rameau peut être très important, surtout dans le cas de présence de cicatrices.

D'autre part, l'évolution des coefficients de régression dans les relations (7) et (14) n'est pas en faveur de l'hypothèse que l'araignée rouge préfère pondre sur les rameaux les plus jeunes. Or toutes les méthodes de prévision actuelles sont basées sur une estimation de la moyenne des populations sur les rameaux de 1 et 2 ans, et donc supposent que ces populations reflètent les populations totales de l'arbre. Certes, cette méthode présente l'avantage du coût : les prélévements des bois de 1 et 2 ans ne causent pas des dommages irréparables à l'arbre, d'autant plus qu'ils peuvent être effectués de façon à concorder avec la taille. Mais l'estimation de la population totale de l'arbre est alors peu précise.

d) Dans quelle mesure peut-on généraliser les résultats du traitement de nos données à d'autres vergers ?

- Le gradient observé sur les effets moyens de la Rangée semble être particulier à notre verger (voir par exemple A-I-[2]).

- Il est étonnant que les effets du facteur Rangée soient significativement non nuls sur les longueurs LLTOT, LL2P, LL1P (G22). Par contre, pour le niveau x face, ceci peut s'expliquer par les effets de la taille.

Ce fait ne peut s'expliquer que par une différence des structures architecturales des haies (présence des peupliers ?), étant donné que tous les prélévements ont été effectués par M. COTTON.

- Mais le plus étonnant est de constater que les classements des effets moyens de la Rangée (et dans une moindre mesure du niveau x face), quand l'analyse est faite sur les longueurs ou sur les populations, sont dans l'ensemble concordantes (même dans le modèle de covariance avec la longueur voir par exemple G15, G17, G22 et G25).

La longueur LONG3 du bois de 3 ans est la seule à échapper à cette constatation (G22). Le tableau T18 montre que c'est aussi le cas où l'apport

des longueurs est très faible (R^2 passe de 0.68 (2) à 0.70 (3) et de 0.78 (6) à 0.79 (7)), mais l'hypothèse de nullité du coefficient de la régression dans le modèle (7) est rejetée au risque de 1%

L'influence de la longueur est donc complexe : les effets de la taille et du fait que les échantillons sont prélevés à la fin de l'hiver ne permettent pas de trouver une explication simple en fonction de la dynamique de population. Le sujet mérite d'être plus approfondi.

e) Afin de poursuivre l'étude, M. FAUVEL a dirigé le prélévement d'un nouvel échantillon dans l'hiver de 81-82 qui devrait enrichir et compléter les résultats trouvés dans diverses directions :

- Relations entre populations sur bois d'une part et populations sur bourses et petites productions d'autre part.

- Relations entre populations sur bois d'âges différents.

- Relations entre populations et caractéristiques architecturales, position dans le verger.

- Relations entre populations et pourcentage d'organes infestés.

BIBLIOGRAPHIE :

[1] *J. BACHACOU - J.P. MASSON - C. MILLIER (1981)*

Manuel de la programmathèque Statistique AMANCE 81
Département de Biométrie - I.N.R.A.

[2] *G.E.P. BOX et G.C. TIAO (1973)*

Bayesian Inference in Statistical analysis

[3] *E. LESQUOY - C. MILLIER - R. TOMASSONE*

La régression (A paraitre aux Ed. MASSON)

[4] *E.S. PEARSON - H.P. HARTLEY (1970)*

Biometrika tables for statisticians - Volume I
3ème édition - Cambridge U.P.

[5] *G.W. SNEDECOR - W.G. COCHRAN (1971) (traduit)*

Méthodes statistiques
A.C.T.A.

P A R T I E B

APPLICATION DE LA THEORIE DES SONDAGES
A UNE STRUCTURE RAMIFIEE

B - APPLICATION DE LA THEORIE DES SONDAGES

Dans cette partie, le problème considéré est assez général, il consiste à :

Estimer la moyenne d'une variable par arbre (ou par branche d'un âge donné) sur une plantation (non nécessairement structurée comme un verger industriel) en tenant compte du fait que pour effectuer la mesure de la variable il est nécessaire de couper le rameau considéré.

Le problème est d'abord formulé mathématiquement, l'application de la théorie des sondages est faite ensuite par étapes en accordant une importance particulière aux dommages causés à l'arbre et à la plantation à tous les niveaux.

Trois hypothèses ont été retenues :

- la répartition de la variable sur les bois de même âge du verger peut ne pas être uniforme (présence de concentration ou d'aggrégats).

- la fonction densité de la variable est absolument continue.

- il y a une bonne relation de régression expliquant la mesure de la variable sur un rameau par celles sur les bois plus jeunes portés par ce rameau (et éventuellement des caractéristiques architecturales des branches).

Autant que possible, nous nous sommes suffi de notations assez répandues dans la théorie des sondages. Mais insuffisantes devant la complexité du problème, d'autres notations ont été définies au fur et à mesure. Tout en essayant de les simplifier au maximum, elles suivent une logique et présentent une homogénéité dans toute la suite :

- les indices i, j et r désignent des rameaux (ou individus)

- les indices h, k et ℓ désignent des âges (ou classes)

- *m, n, v, M, N désignent des effectifs*

- *d'autres notations sont moins rigoureuses, telles que : ∀ij qui signifie pour tous rameaux i et j, et devrait être écrit : ∀i, ∀j.*

B-I - POPULATION

Nous décrivons dans 1) notre population totale $\mathcal{U}$ de tous les rameaux du verger grâce aux hypothèses H1, H2, H3 et H4. Cette structuration est en fait la base de tout ce qui suit et est importante pour comprendre les notations et conclusions faites après.

H1, H2 et H3 ne sont pas vraiment des hypothèses dans le sens classique du terme, puisque toute population ramifiée (fleuves, arbres,...) les vérifie automatiquement, tandis que H4 est une hypothèse qui régularise cette population et nous permet d'aller plus loin.

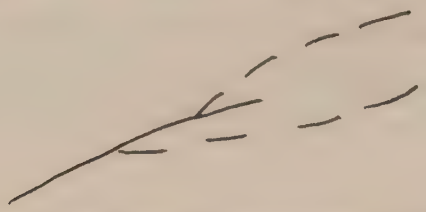
Cependant, du point de vue mathématique, ce travail est valable pour toute population "ramifiée" (dans le sens où elle vérifie H1, H2, H3 et H4) et non seulement pour notre population initiale des rameaux d'un verger.

Dans 2), nous déduisons de ce qui précède une expression claire de notre objectif et fixerons des notations pour la suite.

Pour des raisons de clarté, nous chercherons à utiliser des termes qui se rapprochent le plus possible des termes biologiques (rameaux, branche, arbre...) de façon à avoir toujours présent à l'esprit la structure particulière de notre population et faciliter l'interprétation d'hypothèses ou conclusions faites dans la suite.

En particulier nous précisons dès maintenant le sens donné au mot "rameau" qui désigne un élément quelconque de $\mathcal{U}$.

Déf. : rameau : segment sans ramifications et d'âge constant (correspond à la pousse d'une année)



(voir également remarques R3 et R5).

B-I-1) Considérons la population U de tous les rameaux du verger.

Soit U_k la sous population de tous les rameaux d'âge k de ce verger, et L l'âge de son plus vieil arbre (voir R2).

Si un rameau est d'âge k, et si $k \neq \ell$, alors il n'est pas d'âge ℓ. D'où l'hypothèse (évidente) :

$$\underline{H1} : \begin{cases} U = \bigcup_{k=1\ldots L} U_k \\ \forall k \neq \ell \quad U_k \cap U_\ell = \emptyset \end{cases}$$

Une numérotation de tous les individus de U, à partir de la ramification, est possible. (voir R6)

Nous supposons donc (N_k étant le nombre de rameaux d'âge k) :

$$\underline{H2} : \begin{cases} \text{Pour } k = 1\ldots L \\ \text{tout individu de } U_k \text{ est identifié par un numéro :} \\ U_k = \{1\ldots N_k\},\ N_k \text{ fini et connu} \end{cases}$$

Le nombre N_k de rameaux d'âge k est fini, mais pour k petit (rameaux de 1 an, 2 ans, etc...) N_k est très grand et inconnu.

Notons N le nombre de tous les rameaux du verger. D'après H1, nous avons :

$$N = \sum_{k=1\ldots L} N_k$$

Notre population se distingue par la structure ramifiée que nous traduisons par une relation d'ordre total particulière R sur $\mathcal{U}$.

Nous dirons que deux éléments i et j de $\mathcal{U}$ sont en relation (R(i,j)=1) si i est porté par j.

Montrons que R est une relation d'ordre total :

- *R est réflexive :*

  tout rameau i est porté par lui-même

  $\forall i \in \mathcal{U} \quad R(i,i) = 1$

- *R est antisymétrique :*

  si le rameau i est porté par j, alors j n'est pas porté par i (sauf si i=j)

  $\forall i \neq j \in \mathcal{U} \quad R(i,j) = 1 \implies R(j,i) = 0$

- *R est transitive :*

  si i est porté par j et si j est porté par r, alors i est porté par r

  $\forall i,j,r \in \mathcal{U}$

  $$\left.\begin{array}{l} R(i,j) = 1 \\ R(j,r) = 1 \end{array}\right| \implies R(i,r) = 1$$

En plus, R vérifie les 3 propriétés suivantes :

a) si deux rameaux différents sont de même âge, alors ils ne peuvent être portés l'un par l'autre.

b) tout rameau ne peut porter que des rameaux plus jeunes.

c) pour k et ℓ fixés, $k < \ell$, tout rameau d'âge k est porté par au plus un rameau d'âge ℓ.

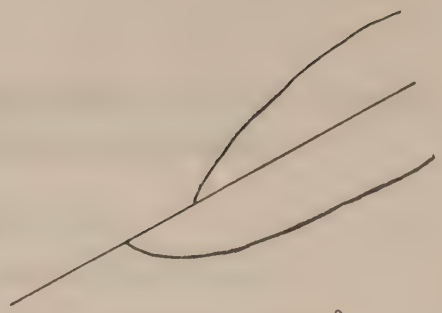
Nous résumons ceci par :

<u>H3</u> : *Il existe une relation d'ordre total R sur $\mathcal{U}$ vérifiant les trois propriétés :*

a) $\forall k, \forall i \neq j \in \mathcal{U}_k \qquad R(i,j)=0$

b) $\forall k < \ell, \forall i \in \mathcal{U}_k, \ \forall j \in \mathcal{U}_\ell \qquad R(j,i)=0$

c) $\forall k < \ell, \forall i \in \mathcal{U}_k$ *il existe au plus* $j \in \mathcal{U}_\ell$ *tel que* $R(i,j)=1$



Déf. : *Nous appelons branche i, notée* W_i, *l'ensemble de tous les rameaux portés par i.*

Donc : $\forall \ell$, $\forall i \in U_\ell$ *(pour tout âge* ℓ, *et toute branche d'âge* ℓ*)*

$$W_i = \{j \in U \text{ tel que } R(j,i) = 1\}$$

D'après H1 et H3-b) on en déduit :

$$W_i = \bigcup_{k \leqslant \ell} \{j \in U_k \text{ tel que } R(j,i) = 1\}$$

Soient i et j deux éléments de U_ℓ, *et supposons :*

$$\exists\, r \in (W_i \cap W_j)$$

alors, par définition de W_i :

$$R(r,i) = R(r,j) = 1$$

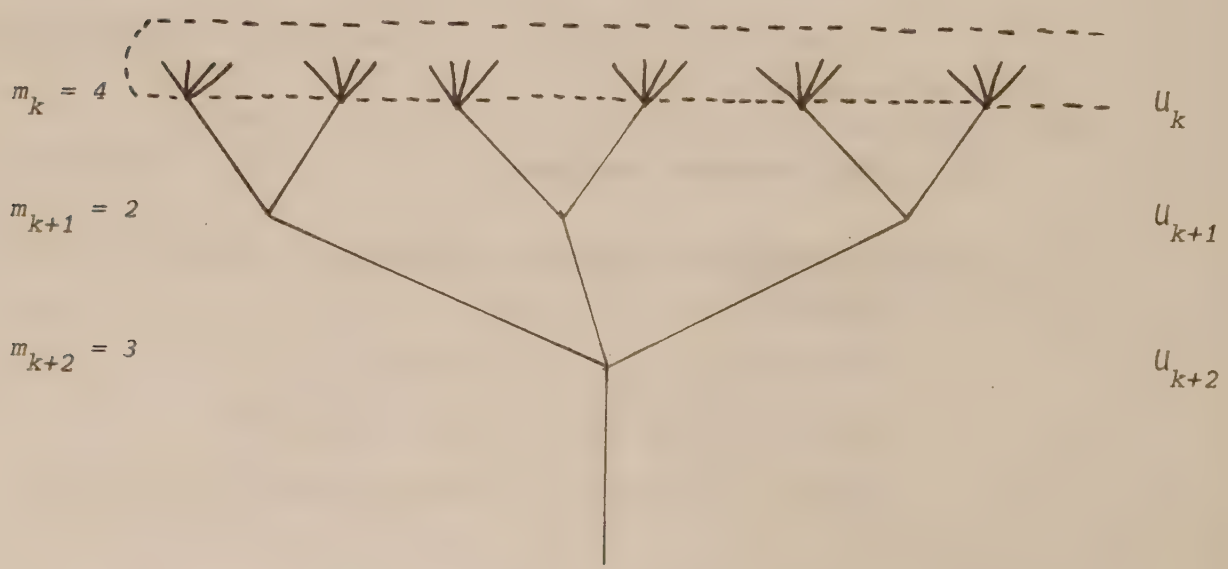
et d'après H3-c), on a nécessairement i=j.

d'où

C1 : $$\boxed{\forall \ell \text{ , } \forall i \neq j \in U_\ell \qquad W_i \cap W_j = \emptyset}$$

Nous supposerons en plus que l'hypothèse suivante est vérifiée. H4 a pour but de régulariser notre population.

H4 : $$\boxed{\forall k < L \text{ , } \forall i \in U_{k+1} \quad card(W_i \cap U_k) = m_k \geqslant 1}$$



Posons $m_L = 1$

$$M_k = \prod_{\ell=k\ldots L} m_\ell$$

et $\forall\, k \leq \ell \qquad M_{k\ell} = M_k / M_\ell$

Nous avons : $\forall\, k < L \qquad m_k = M_{k,k+1}$

$\forall\, h \leq k \leq \ell \quad M_{hk} \cdot M_{k\ell} = M_{h\ell}$

H4 signifie que tout rameau d'âge k+1 porte le même nombre de rameaux d'âge k.

Nous en déduisons que tout rameau d'âge k est porté par un unique rameau d'âge ℓ, pour tout $\ell > k$, et que tous les arbres ont le même âge L.

C2 : $\forall\, k \leq \ell\ ,\ \forall\, j \in U_k\ ,\ \exists!\ i \in U_\ell$ *tel que* $j \in W_i$

C3 : $U = \bigcup_{k=1\ldots L} U_k = \bigcup_{i \in U_L} W_i$

C3 découle du fait que tous les arbres ont le même âge et donc que tout rameau j de U est porté par un arbre (ou encore par un rameau d'âge L).

Ainsi, nous avons défini une deuxième partition de U , correspondant aux arbres dans le verger.

A partir de H1, H2, H3 et C2 nous déduisons C4 et C5 :

C4 :

a) $\forall\, k \leq \ell \qquad U_k \subset \bigcup_{i \in U_\ell} W_i$

b) $\forall\, h \leq \ell \qquad \forall\, i \in U_\ell\ ,\ \forall k$ *tel que :* $h \leq k \leq \ell$

$$W_i \cap U_h = \Big(\bigcup_{j \in W_i \cap U_k} W_j \Big) \cap U_h$$

Dem :

a) $\forall\, j \in U_k\ ,\ \exists\, i \in U_\ell$ *tel que* $j \in W_i$ *(C2)*

b) *Soit* $j \in W_i \cap U_h$ *, d'après C2 :*

$$\exists\ r \in U_k \text{ tel que } j \in W_r \cap U_h$$

On a donc :

$$R(j,i) = 1$$
$$R(j,r) = 1$$

D'après C2, $\exists\ i' \in U_\ell$ *tel que* $R(r,i') = 1$

Appliquons la transitivité, nous avons :

$$R(j,i) = R(j,i') = 1$$

et d'après H3-c) : $i = i'$

Donc :

$$\forall\ j \in W_i \cap U_h\ ,\ \exists\ r \in W_i \cap U_k \text{ tel que :}$$
$$j \in W_r \cap U_h$$

Soit : $W_i \cap U_h \subset \left(\bigcup_{r \in W_i \cap U_k} W_r \right) \cap U_h$

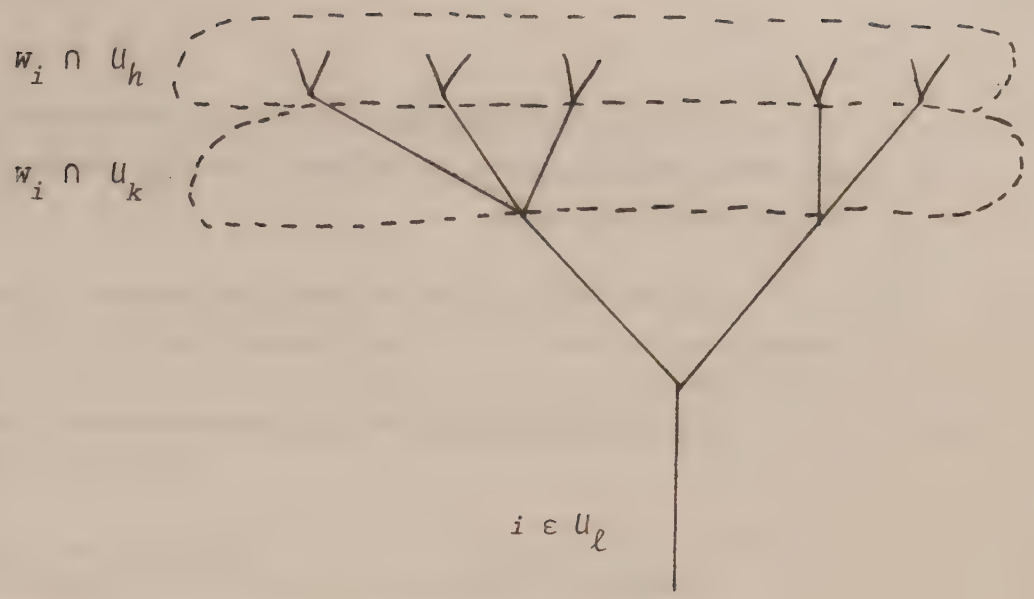
D'autre part, d'après C2, nous avons :

$$W_i \cap U_h \supset \left(\bigcup_{r \in W_i \cap U_k} W_r \right) \cap U_h$$

puisque $h \leq \ell$

D'où le résultat.

C4-b) signifie que l'ensemble des rameaux d'âge h portés par le rameau i d'âge ℓ *est égal à l'ensemble des rameaux d'âge h portés par l'ensemble des rameaux d'âge k portés par i, pour tout h, k et* ℓ *tels que* $h \leq k \leq \ell$.



<u>C5</u> :

$\forall k < L \quad , \quad \forall i \neq i' \in U_k$

un seul des deux cas suivants est possible :

a) soit $\exists\, j \neq j' \in U_L$ *tels que :*

$i \in W_j$ *et* $i' \in W_{j'}$,

b) soit $\exists!\, \ell \quad , \quad k \leq \ell < L$ *tel que :*

$\exists\, j \in U_{\ell+1} \quad , \quad \exists\, r \neq r' \in W_j \cap U_\ell$

avec $i \in W_r$ *et* $i' \in W_{r'}$,

C5 signifie que si i et i' sont des rameaux différents d'âge k, k < L, alors :

a) soit ils sont portés par des arbres différents j et j'

b) soit ils sont portés par le même arbre, et alors il existe un unique ℓ, tels que i et i' soient portés par des rameaux différents d'âge ℓ portés eux-mêmes par le même rameau d'âge $\ell + 1$.

De H1, H2, H3 et H4, nous déduisons :

<u>C6</u> :

$\forall\, k \leq \ell$

a) $N_k = M_{k\ell} \cdot N_\ell$

b) $\forall\, i \in U_\ell \quad card(W_i \cap U_k) = M_{k\ell} \geq 1$

Un cas particulier de C6-a) est :

$$\forall\, k \quad N_k = M_k \cdot N_L$$

Ainsi, il suffira de connaitre N_L *: la taille de la population la plus âgée, et* $\{m_k ; k = 1...L-1\}$ *pour que* $\{N_k ; k = 1...L\}$ *soient connus et l'hypothèse H2 vérifiée.*

C6-b) signifie que le nombre de rameaux d'âge k portés par le même rameau d'âge ℓ*, avec* $\ell > k$*, est constant, égal à* $M_{k\ell}$*, et est non nul.*

M_k *est alors le nombre de rameaux d'âge k portés par tout arbre.*

Dem :

⁂*Montrons d'abord C6-b) par récurrence sur* ℓ

- *pour* $\ell = k+1$ *, d'après H4 :*

  $\forall\, j \in U_{k+1} \quad card(W_j \cap U_k) = m_k = M_{k,k+1}$

- *supposons que* $\forall\, r \in U_{\ell-1}$

  $card(W_r \cap U_k) = M_{k,\ell-1}$

- *appliquons C4-b) :*

  $$\forall\, i \in U_\ell \quad W_i \cap U_k = \bigcup_{r \in W_i \cap U_{\ell-1}} W_r \cap U_k$$

  d'où (C1) : $Card(W_i \cap U_k) = Card(W_i \cap U_{\ell-1}) \cdot Card(W_r \cap U_k)$

  $= M_{\ell-1,\ell} \cdot M_{k,\ell-1} = M_{k\ell}$

⁂*Appliquons C4-a) :*

$$U_k = \bigcup_{i \in U_\ell} (W_i \cap U_k)$$

Soit $N_k = N_\ell \cdot M_{k\ell}$

B-I-2) Dans la suite, nous désignerons par y_{ki} la mesure de la caractéristique y sur le rameau i d'âge k. y_k désignera le vecteur de R^{N_k} de composante y_{ki} , $i = 1...N_k$ et :

$$\bar{y}_k = \frac{1}{N_k} \sum_{i=1...N_k} y_{ki}$$

la moyenne sur tout le verger des y_{ki} sur tous les rameaux d'âge k.

Nous désignerons par y_m la moyenne par arbre sur tout le verger de la caractéristique y.

Sous les hypothèses H1...H4, y_m s'écrit :

$$y_m = \frac{1}{N_L} \sum_{i \in U_L} \left[y_{Li} + \sum_{k=1...L-1} \sum_{j \in W_i \cap U_k} y_{kj} \right]$$

$$y_m = \frac{1}{N_L} \sum_{k=1...L} \sum_{i \in U_L} \sum_{j \in W_i \cap U_k} y_{kj}$$

$$= \frac{1}{N_L} \sum_{k=1...L} \sum_{j \in U_k} y_{kj} \qquad (C4\text{-}a)$$

$$= \sum_{k=1...L} \frac{N_k}{N_L} \bar{y}_k$$

$$= \sum_{k=1...L} M_k \bar{y}_k \qquad (C6\text{-}a)$$

Nous supposerons que l'hypothèse H5 est vérifiée :

H5 : [Toute mesure est effectuée sans erreur

R1 : Notons $\bar{y}$ la valeur moyenne par rameau sur tout le verger que prend la caractéristique y sur U.

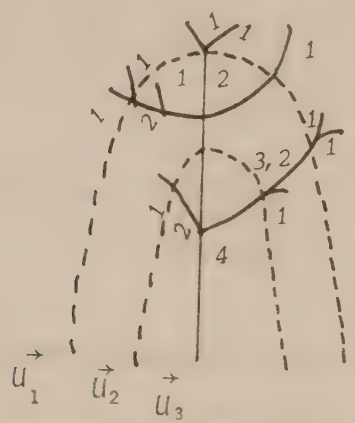
$$\bar{y} = \frac{1}{N} \sum_{k=1...L} \sum_{i \in U_k} y_{ki}$$

$$\text{avec } N = \sum_{k=1...L} N_k = N_L \sum_{k=1...L} M_k \qquad (C6\text{-}a)$$

$$\text{alors } y_m = \bar{y} \cdot \sum_{k=1...L} M_k$$

R2 : *L n'est pas nécessairement l'âge du plus vieil arbre du verger. Nous pouvons restreindre notre étude aux branches d'âge k, k donné.* U_ℓ *est alors défini comme l'ensemble des rameaux d'âge* ℓ *portés par les branches d'âge k et* $U = \bigcup_{\ell \leq K} U_\ell$

R3 : *La définition des rameaux et la première partition de la population initiale ont été faites à partir de l'âge.*
Ce que nous entendons par âge ne correspond pas nécessairement à la définition biologique du terme.
Selon la variable considérée, il peut être défini à partir de critères structuraux (distance par rapport au tronc, ou à la surface foliaire, diamètre, etc...) ou biologique (âge...).
S'il s'agit par exemple d'estimer le nombre de parasites se propageant à partir de la surface foliaire vers le tronc, il serait plus judicieux de considérer la partition suivante (l'âge biologique est indiqué pour chaque rameau) :



Pour nos données, oeufs d'hiver de l'araignée rouge Panonychus Ulmi déposés sur un verger de pommiers, nous considérons l'âge biologique des rameaux.

R4 : *La démarche suivie jusqu'à maintenant peut s'appliquer à toute population ramifiée (fleuve, etc...)*

R5 : *La définition du mot "rameau" est en général assez floue. Pour les arbres ramifiés à axes végétatifs tous équivalents et orthotropes, ce que nous entendons par "rameau" correspond à l' "article" tel que le définissent HALLE et OLDEMANN :*

> *"Un article est un axe dont le méristème édificateur parcourt la totalité de la séquence de différenciations propre à l'espèce considérée, depuis son origine jusqu'à la sexualité"*

R6 : L'hypothèse H2 est très importante pour la suite. Elle suppose que notre population ne corresponde à aucun des 2 cas suivants :

- population des poissons dans un lac : ni l'effectif N de la population est connu, ni les individus sont identifiables.

- une urne contenant les boules non marquées.

Ici l'effectif N peut être connu, mais les éléments sont non identifiables.

En effet, nous pouvons utiliser la propriété que tout rameau est porté par un arbre, et peut être porté par des rameaux plus âgés.

Nous proposons la procédure suivante, utilisée pour les données 81, mais d'une façon plus compliquée :

- à chaque arbres, nous associons un numéro i_L , $i_L=1...N_L$

- pour k=L-1...1, à chaque rameau d'âge k nous associons $i_L,...,i_{k+1}$, i_k avec :

  . i_L le numéro de l'arbre porteur

  . i_ℓ ($\ell = L \ldots k+1$) le numéro du rameau d'âge ℓ porteur du rameau considéré.

i_ℓ est nul si aucun rameau d'âge ℓ ne porte le rameau considéré.

- pour l'attribution du numéro i_k , nous proposons la règle suivante :

Considérons deux rameaux j et j' portés par les mêmes rameaux $i_L\ i_{L-1} \ldots i_{k+1}$. Le numéro $i_k(j)$ attribué à j est supérieur à $i_k(j')$ attribué à j' si l'origine de j est plus éloignée que celle de j' de la base du rameau porteur. Dans le cas des ramifications verticillées, on peut adopter une règle de classification supplémentaire, basée sur l'angle entre : le plan défini par les 2 rameaux porteurs les plus jeunes, et le plan défini par le rameau considéré et le rameau porteur, et étant donné un sens de rotation (bonhomme d'Ampère).

Si H4 est vérifiée, il suffit alors d'identifier les arbres pour que tous les individus le soient.

B I B L I O G R A P H I E :

[1] *F. HALLE & R.A.A. OLDEMAN (1970)*

Essai sur l'architecture et la dynamique de croissance des arbres tropicaux.

Masson & Cie

B-II - PLAN DE SONDAGE

Comment déterminer une méthode pour prélever un échantillon d'une population ramifiée ? Ou plutôt, quelles sont les propriétés que vérifie un plan de sondage sur notre population ?

Dans 1), nous fixons les notations et nous limitons aux plans non ordonnés, sans remise et non informatifs.

Nous nous intéressons particulièrement aux probabilités d'inclusion d'ordre 1 et 2 ; ce sont elles seules qui interviendront lors du calcul des espérances et variances de toute fonction des valeurs mesurées sur l'échantillon prélevé $\{y_i \; ; \; i \in S\}$.

Un 1er théorème est alors énoncé (1 - g) dû à GODAMBE (1955) et à HANURAV (1962).

Dans 2), nous énonçons une propriété fondamentale de notre plan : nous ne prélevons pas des rameaux, mais des branches (H9).

La "contrainte d'inclusion" H9 est faite pour une raison de coût évidente : causer le moins de dommages à un verger et donc tirer le maximum d'informations de toute branche coupée.

Nous en déduisons alors des premières conséquences sur les probabilités d'inclusion d'ordre 1 et 2.

Dans 3), nous utilisons nos connaissances a priori sur la variable à estimer : nombre d'oeufs d'hiver de P. ULMI sur un verger de pommiers, pour faire l'hypothèse H1O sur les probabilités d'inclusion d'ordre 1. Nous en déduisons les probabilités d'inclusion d'ordre 1 de tout rameau du verger en fonction de celles des arbres porteurs.

Dans 4), nous nous limitons aux plans de sondages à taille fixe. Dans 4-c), nous résumons les résultats donnant les probabilités d'inclusion d'ordre 2 en fonction de celles des arbres porteurs.

Parallélement, des paramètres β_k , γ_k et δ_k associés à chacun des âges sont définis. Leurs propriétés sont résumées dans C13 et C22.

En fait, il suffit d'en connaitre 2. Nous retenons β_k et γ_k , k=1...L. β_k (resp. γ_k) est directement lié à la taille (resp. la dispersion) de l'échantillon s_k , mais varie dans un sens inverse.

Dans 5) nous indiquons brièvement les liens entre probabilités d'inclusion et plans de sondages et comment reconstituer les uns à partir des autres.

B-II-1) Notations et définitions générales - Introduction

a) Pour tout k, k=1...L

Soit s_k l'échantillon des rameaux d'âge k prélevés et n_k son cardinal

$$s_k \subset U_k \qquad n_k = Card(s_k)$$

A chaque échantillon s_k nous faisons correspondre l'ensemble des données :

$$d_k = \{(i, x_{ki}, y_{ki}),\ i \in s_k\}$$

Chaque donnée est un $(n_x + n_y + 1)$ - uplet où :

- i : est le numéro du rameau prélevé
- x_{ki} : le vecteur ligne des mesures des n_x variables auxiliaires sur l'individu i de U_k
- y_{ki} : le vecteur ligne des mesures des n_y variables inconnues qui nous intéressent sur l'individu i de U_k

Notons s l'échantillon total prélevé.

D'après H1, s peut se décomposer de façon unique en des sous échantillons s_k , k=1...L, tel que :

$$s = \bigcup_{k=1...L} s_k$$

A s nous associons :

$$d = \bigcup_{k=1...L} d_k$$

Notons : $\forall k, \bar{s}_k = U_k - s_k$

$$\bar{s} = U - s$$

b) Une hypothèse, justifiée en III-2 , est que l'ordre dans lequel les différents rameaux d'un échantillon s_k sont prélevés n'a aucune importance.

H6 : [Pour $k=1...L$, s_k est non ordonné.
Il s'ensuit que s est aussi non ordonné.

H6 signifie que s_k peut s'écrire :

$$s_k = \{i_1 , \ldots , i_{n_k}\} = \{i_{\phi(1)} , \ldots , i_{\phi(n_k)}\}$$

où ϕ est une permutation quelconque de $\{1 \ldots n_k\}$ et i_j désigne le jème rameau prélevé.

Nous noterons : $s_k = \{i_1 , \ldots , i_{n_k}\}$

avec : $i_1 \leqslant i_2 \leqslant \ldots \leqslant i_{n_k}$

c) Pour prélever un rameau, il est nécessaire de le couper. Tous les rameaux prélevés sont donc différents, d'où :

H7 :[Pour $k=1,\ldots,L$, s_k est sans remise

Donc s est aussi sans remise. (voir III-2)

En particulier nous avons : $\forall k$, $n_k \leqslant N_k$

d) Notons :

$S_k = \{s_k \subset U_k$ tel que : s_k non ordonné, sans remise$\}$

a_k : la σ algèbre des parties de S_k

et soit S_k la variable aléatoire : échantillon prélevé de la population U_k.

(S_k , a_k) est un espace probabilisable.

Soit $P(S_k = s_k)$ la probabilité que l'échantillon prélevé de U_k soit s_k, définie pour tout s_k de S_k.

Nous appelons "Plan de sondage" le triplet (S_k , a_k , P), et par abus : P_k

Nous supposerons que pour tout âge k, la probabilité de prélever un échantillon s_k est indépendante des mesures $\{(y_{ki}), i \in s_k\}$

H8 : [Pour $k=1...L$, P_k est non informatif.

L'hypothèse H8 répond surtout à un esprit pratique, car sinon, cela suppose qu'après chaque prélévement, il faut mesurer les variables qui nous intéressent, et selon les résultats déjà observés, décider le prochain individu à prélever.

Par contre, H8 ne concerne pas tout ce qu'on peut avoir comme information a priori (stratification, variables auxiliaires x), qui, elle, peut être incluse au niveau de la détermination du plan de sondage.

Pour toute la suite, et pour toute fonction $F(S_k)$, *nous noterons :*

$$E_{p_k} F(S_k) = \sum_{s_k \in S_k} F(s_k) \cdot P(S_k = s_k)$$

e) Définissons les variables aléatoires :

- $\forall i \in U_k$

$$1_i(S_k) = \begin{cases} 1 & si \quad i \in S_k \\ 0 & si \quad i \notin S_k \end{cases}$$

- $\nu_k = \sum_{i \in U_k} 1_i(S_k)$

ν_k *est la taille aléatoire de l'échantillon prélevé de* U_k

et notons les probabilités d'inclusions :

- $\alpha_i = P(i \in S_k) = E_{p_k} 1_i(S_k) \quad \forall i \in U_k$

- $\forall i \neq j \in U_k$

$$\alpha_{ij} = P(i,j \in S_k) = E_{p_k} 1_i(S_k)\ 1_j(S_k)$$

Nous dirons que le plan de sondage P_k *est à taille fixe* n_k *si :*

$$P(\nu_k = n_k) = 1$$

Sinon, dans le cas général, on note :

$$n_k = E_{p_k} \nu_k$$

n_k *est la taille moyenne de* S_k

f) Pour la population totale U, nous définissons de même :

$$S = \{s = \bigcup_{k=1...L} s_k \text{ tel que } \forall k, s_k \in S_k\}$$

ainsi que a *, la v.a . S et le plan de sondage P.*

D'après H1, $\forall i \in U$, $\exists! k$ *tel que* $i \in U_k$*, nous avons alors :*

$$1_i(S) = \sum_{h=1...L} 1_i(S_h) = 1_i(S_k)$$

$$\alpha_i = P(i \in S) = \sum_{h=1...L} P(i \in S_h) = P(i \in S_k)$$

Par contre, pour les probabilités d'inclusion d'ordre ≥ 2*, et pour les plans de sondage, les relations sont plus complexes.*

g) Dans la suite, nous nous intéressons aux probabilités d'inclusion d'ordre 1 et 2, ce sont les seules qui interviennent pour le calcul des espérances et variances des fonctions linéaires de S.

Pour toute population U finie, d'effectif connu N, dont tous les individus sont répertoriés, et pour tout plan de sondage P non informatif, nous avons le théorème suivant :

Th1 :

a) $\sum_{i \in U} \alpha_i = E_p \nu$

$\forall i \in U \quad \sum_{\substack{j \in U \\ j \neq i}} \alpha_{ij} = E_p[1_i(S)\nu] - \alpha_i$

b) si P est de taille fixe n, alors :

$\sum_{i \in U} \alpha_i = n$

$\forall i \in U \quad \sum_{\substack{j \in U \\ j \neq i}} \alpha_{ij} = \alpha_i(n - 1)$

Dans les deux cas :

- *population totale U de tous les rameaux du verger*
- *sous populations U_k des rameaux d'âge k, k=1...L*

nous pouvons appliquer le Th1. Nous inclurons les résultats de l'analyse des données et les particularités de notre population pour déterminer complétement les probabilités d'inclusion d'ordre 1 et 2.

R7 : Le Th1 est valable même si les hypothèses H6 et H7 ne sont pas vérifiées.
Dans le cas où l'échantillon est avec remise, ν est sa taille effective : le nombre d'éléments distincts qui ont été prélevés.

R8 : Pour simplifier les notations, nous avons utilisé W_i, α_i ... sans préciser l'âge. Dans le cas d'un risque de confusion, celui-ci est précisé juste avant.

R9 : L'hypothèse H8 n'exclut pas l'échantillonnage séquentiel.
WALTER a appliqué cette technique au cas des insectes dans une forêt, dans le cas où les tailles des populations sont très grandes.
Connaissant un certain seuil, et la valeur des 2 risques de 1ère et 2ème espèce, on détermine 2 valeurs m_1 et m_2. On décide :

- soit d'effectuer le traitement si $y_m > m_2$
- de ne pas l'effectuer si $y_m < m_1$
- sinon de prélever encore des individus et de reprendre dès le début.

Si les individus à prélever ne sont pas déterminés en fonction des mesures déjà faites, l'échantillonnage séquentiel revient alors à une séquence d'échantillons emboités de tailles croissantes.

Pour nous, on cherche à estimer la valeur y_m et non à prendre une décision : traiter ou non, qui suppose la prédétermination d'un seuil.

B-II-2) Contrainte d'inclusion

Le plan de sondage P se distingue par la propriété suivante :

Pour prélever un rameau i, il est nécessaire de le couper, et donc en fait, nous prélevons toute la branche i.

$$\underline{H9} : \begin{cases} \forall k \geq 2 \\ i \in s_k \Rightarrow \forall h < k \ , \ \forall j \in W_i \cap U_h \ , \ j \in s_h \end{cases}$$

H9 s'écrit aussi :

$$\forall k \geq 2 \ , \forall i \in U_k \ , \forall h < k \ , \forall j \in W_i \cap U_h$$

$$P(j \in S_h \ / \ i \in S_k) = 1$$

ou encore :

$$P(i \in S_k \ / \ j \notin S_h) = 0$$

Nous en déduisons C7, C8, C9, C10 qui établissent les relations induites par H9 pour les probabilités d'inclusion d'ordre 1 et 2.

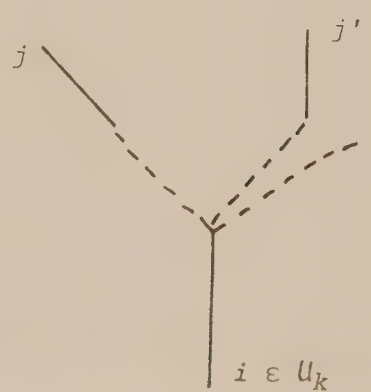
$$\underline{C7} : \begin{cases} \forall h < k \ , \forall i \in U_k \ , \forall j \in W_i \cap U_h \\ \alpha_j = \alpha_i + (1 - \alpha_i) \ P(j \in S_h \ / \ i \notin S_k) \geq \alpha_i \end{cases}$$

En effet :

$$\alpha_j = P(j \in S_h) = P(j \in S_h \cap i \in S_k) + P(j \in S_h \cap i \notin S_k)$$

$$= \alpha_i \ P(j \in S_h \ / \ i \in S_k) + (1 - \alpha_i) \ P(j \in S_h \ / \ i \notin S_k)$$

Pour C8, j et j' sont deux rameaux d'âge h portés par le même rameau i d'âge k



$$\underline{C8} : \begin{cases} \forall h < k \ , \ \forall i \in U_k \ , \ \forall j \neq j' \in W_i \cap U_h \\ \alpha_{jj'} = \alpha_i + (1 - \alpha_i) \ P(jj' \in S_h \ / \ i \notin S_k) \geqslant \alpha_i \end{cases}$$

Pour C9 et C10, j et j' sont des rameaux d'âge h portés par des rameaux distincts i et i' d'âge k

$$\underline{C9} : \begin{cases} \forall h < k \ , \ \forall i \neq i' \in U_k \ , \ \forall j \in W_i \cap U_h \ , \ \forall j' \in W_{i'} \cap U_h \\ \alpha_{jj'} = \alpha_{ii'} + (\alpha_i - \alpha_{ii'}) \ P(jj' \in S_h \ / \ i \in S_k \ , \ i' \notin S_k) \\ \qquad + (\alpha_{i'} - \alpha_{ii'}) \ P(jj' \in S_h \ / \ i \notin S_k \ , \ i' \in S_k) \\ \qquad + (1 - \alpha_i - \alpha_{i'} + \alpha_{ii'}) \ P(jj' \in S_h \ / \ i \ i' \notin S_k) \geqslant \alpha_{ii'} \end{cases}$$

En effet :

$$P(i \in S_k \ , \ i' \notin S_k) = \alpha_i - \alpha_{ii'} \geq 0$$

$$P(ii' \notin S_k) = 1 - P(i \in S_k \ , \ i' \notin S_k) - P(i \notin S_k \ , \ i' \in S_k) - P(ii' \in S_k)$$

$$= 1 - \alpha_i - \alpha_{i'} + \alpha_{ii'} \geq 0$$

Si $h=k$ *, nous avons :*

$$0 = P(ii' \in S_k \ / \ i \in S_k \ , \ i' \notin S_k)$$

$$= P(ii' \in S_k \ / \ ii' \notin S_k)$$

C10 : $\forall h \leq k \ , \ \forall i \neq i' \in U_k \ , \ \forall j \in W_i \cap U_h \ , \ \forall j' \in W_{i'} \cap U_h$

$$P(jj' \in S_h \ / \ i \notin S_k \ , \ i' \in S_k) = \frac{1}{\alpha_{i'} - \alpha_{ii'}} \left(\alpha_{i'} \, P(j \in S_h \ / \ i' \in S_k) - \alpha_{ii'} \right)$$

Dem :

$$P(jj' \in S_h \ / \ i \notin S_k \ , \ i' \in S_k)$$

$$= P(j \in S_h \ / \ i \notin S_k \ , \ i' \in S_k) - P(j \in S_h \ , \ j' \notin S_h \ / \ i \notin S_k \ , \ i' \in S_k)$$

$$= P(j \in S_h \ / \ i \notin S_k \ , \ i' \in S_k)$$

$$= \frac{P(j \in S_h \ , \ i' \in S_k) - P(j \in S_h \ , \ ii' \in S_k)}{P(i \notin S_k \ , \ i' \in S_k)}$$

$$= \frac{1}{\alpha_{i'} - \alpha_{ii'}} \left[\alpha_{i'} \, P(j \in S_h / i' \in S_k) - \alpha_{ii'} \, P(j \in S_h / ii' \in S_k) \right]$$

$$= \frac{1}{\alpha_{i'} - \alpha_{ii'}} \left[\alpha_{i'} \, P(j \in S_h / i' \in S_k) - \alpha_{ii'} \right]$$

B-II-3) Probabilités d'inclusion d'ordre 1

a) Pour toute population U vérifiant H2, et pour tout plan de sondage non informatif (H8), donc en particulier pour notre population totale U, et pour chacune des sous populations $U_1 \ldots U_L$, la détermination des probabilités d'inclusion du 1er ordre à partir de nos connaissances a priori sur la variable observée y se résume principalement en 3 possibilités :

a1 : on n'utilise aucune information a priori ; tous les individus sont alors équiprobables :

$$\forall i \in U \qquad \alpha_i = E_p \nu / N$$

$E_p \nu$ (la taille moyenne de l'échantillon) est supposée connue.

a2 : nous pouvons subdiviser U en des sous populations homogènes (faibles variances intra pour la variable y considérée). Les α_i sont prises alors constantes à l'intérieur de chacune d'elles.

a3 : il existe une variable auxiliaire x positive, corrélée fortement avec y et telle que :

$$\forall i \in U \qquad x_i \cdot E_p \nu < N \bar{x}$$

on prend alors :

$$\forall i \in U \qquad \alpha_i = E_p \nu \cdot x_i / N \bar{x}$$

Ces 3 cas correspondent aux possibilités d'affecter un poids à chaque individu :

- tous les individus ont le même poids
- à chaque individu on affecte un poids spécifique.

Dans le cas a3, si la variable auxiliaire est qualitative nous retrouvons le cas a2.

Quand aux conditions sur les valeurs prises par x, elles se justifient par le fait que :

$$\forall i \in U \qquad 0 \leqslant \alpha_i \leqslant 1$$

et par $\sum_{i \in U} \alpha_i = E_p \nu$ (Th1)

Ces conditions ne sont nullement restrictives. Dans le cas où :

$$i \in U \text{ tel que } x_i \cdot E_p \nu > N \bar{x}$$

Il suffit d'appliquer la translation : $x \to x + a$ avec :

$$a > (E_p \nu \cdot \text{Sup } x_i - N \bar{x}) / (N - E_p \nu)$$

b) Quels cas retenir à partir des résultats des traitements des données (A-II) pour chacune des populations U_1, U_2, U_3 ?

. Parmi les caractéristiques architecturales de la branche seule la longueur semble convenir. Il est très probable d'après la bibliographie que d'autres caractéristiques soient intéressantes. tels que le nombre d'obstacles ou de caches potentielles telles que les définissent GASSER et KLAY, ou tout simplement le nombre de bourses.

Considérer ces caractéristiques comme variables auxiliaires sous entend que nous pouvons les supposer connues sur tout le verger quitte à effectuer un large pré-sondage. Or les mesures de telles variables n'est pas aisée, le coût du pré-sondage est assez élevé et l'apport des longueurs relativement faible en comparaison avec l'apport de la stratification (A-II-T18).

Cependant, nous pouvons avoir des variables auxiliaires très intéressantes dans d'autres problèmes (Biomasse...)

Le cas a3 n'est donc pas retenu à ce stade. L'utilisation des longueurs sera revue dans l'estimation.

. A partir des résultats de l'analyse de variance (A-II-6), nous pouvons supposer que les probabilités d'inclusion d'ordre 1 sont constantes pour des rameaux de même âge, de la même rangée et au même niveau x face. Si le problème est d'estimer la moyenne de la variable sur les branches d'âge L ( ) et non sur les arbres, et si L est bien inférieur à l'âge des arbres, alors nous pouvons supposer :

$$H10 : \begin{cases} \forall\, i \in U_L \,,\ \forall k \leq L,\ \forall j \neq j' \in W_i \cap U_k \\ \alpha_j = \alpha_{j'} > 0 \end{cases}$$

l'hypothèse que $\alpha_j > 0$ *sera justifiée dans Th3 et signifie que tout rameau de la population totale* U *a une probabilité non nulle d'être prélevé.*

Dans la suite, nous considérons que cette hypothèse est vérifiée même si le but est d'estimer la population moyenne par arbre.

c) De C7 et H10, nous déduisons :

$$\forall h \leq k\ ,\ \forall i \in U_k\ ;\ \forall j \in W_i \cap U_h$$

$$P(j \notin S_h \,/\, i \notin S_k) \text{ est constante}$$

Notons :

. $\forall h \leq k\ ,\ \forall i \in U_k\ ,\ \forall j \in W_i \cap U_h$

$$\beta_{hk} = P(j \notin S_h \,/\, i \notin S_k)$$

. $\forall k \leq L \quad \beta_k = \beta_{kL}$

<u>C11</u> :

. $\forall h \leq k \leq \ell$

$$\beta_{h\ell} = \beta_{hk} \cdot \beta_{k\ell}$$

. $1 = \beta_L \geq \beta_{L-1} \geq \ldots \geq \beta_1 \geq 0$

En effet : soit j_h *un rameau d'âge h porté par le rameau* j_k *d'âge k, lui même porté par* j_ℓ *d'âge* ℓ

De H9, nous avons :

$$0 = P(j_h \notin S_h \cap j_k \in S_k) = P(j_h \notin S_h \cap j_\ell \in S_\ell)$$

$$= P(j_k \notin S_k \cap j_\ell \in S_\ell)$$

d'où :

$$P(j_h \notin S_h) = P(j_h \notin S_h \cap j_k \notin S_k)$$
$$= P(j_h \notin S_h \cap j_\ell \notin S_\ell)$$
$$P(j_k \notin S_k) = P(j_k \notin S_k \cap j_\ell \notin S_\ell)$$
$$\beta_{h\ell} = P(j_h \notin S_h / j_\ell \notin S_\ell)$$
$$= P(j_h \notin S_h) / P(j_\ell \notin S_\ell)$$
$$= P(j_h \notin S_h / j_k \notin S_k) \cdot P(j_k \notin S_k / j_\ell \notin S_\ell)$$
$$= \beta_{hk} \cdot \beta_{k\ell}$$

Avec ces notations, et sous l'hypothèse H10, C7 devient :

$$\underline{C12} : \left[\begin{array}{l} \forall h < k \ , \ \forall i \in U_k \ , \ \forall j \in W_i \cap U_h \\ 1 - \alpha_j = \beta_{hk} \cdot (1 - \alpha_i) \end{array}\right.$$

Appliquons alors le Th1 :

$$\sum_{j \in U_h} (1 - \alpha_j) = N_h - E_{p_h} \nu_h$$
$$= \beta_{hk} \sum_{i \in U_k} \sum_{j \in W_i \cap U_h} (1 - \alpha_i) \qquad (c4\text{-}a)$$
$$= \beta_{hk} \cdot M_{hk} \sum_{i \in U_k} (1 - \alpha_i)$$
$$= \beta_{hk} \cdot M_{hk} \cdot (N_k - E_{p_k} \nu_k)$$

$$\underline{C13} : \left[\forall h < k \qquad N_h - E_{p_h} \nu_h = \beta_{hk} \cdot M_{hk} \ (N_k - E_{p_k} \nu_k)\right.$$

En particulier :

- la taille de l'échantillon prélevé de toute population U_k , k=1...L-1, s'exprime en fonction de celle de l'échantillon sur U_L (arbres) par :

$$N_k - E_{p_k} \nu_k = \beta_k \cdot M_k \cdot (N_L - E_{p_L} \nu_L)$$

- pour tout rameau j de U, nous savons déterminer son âge k et l'arbre porteur i alors :

$$1 - \alpha_j = \beta_k \cdot (1 - \alpha_i)$$

B-II-4) Probabilités d'inclusion d'ordre 2

a) Pour k=1...L-1, nous allons déterminer les probabilités d'inclusion d'ordre 2 de la population U_k en fonction de celles de la population U_{k+1}.

Nous supposons que les plans de sondages vérifient les deux hypothèses :

H11 : *les plans de sondages sont à taille fixe*

H12 :
- *Si* $m_k \geq 2$ *alors* $\forall i \in U_{k+1}$, $\forall j \neq j' \in W_i \cap U_k$

  $P(jj' \in S_k \,/\, i \notin S_{k+1}) = b_k$

  Si $m_k = 1$, *nous posons* $b_k = 0$
- $\forall i \neq i' \in U_{k+1}$, $\forall j \in W_i \cap U_k$, $\forall j' \in W_{i'} \cap U_k$

  $P(jj' \in S_k \,/\, i \in S_{k+1} \,,\, i' \notin S_{k+1}) = c_k$

  $P(jj' \in S_k \,/\, ii' \notin S_{k+1}) = d_k$

Pour homogénéiser les notations, posons :

$\forall k < L$, $\forall i \in U_{k+1}$, $\forall j \in W_i \cap U_k$

$P(j \in S_k \,/\, i \notin S_{k+1}) = a_k$

Nous avons : $a_k = 1 - \beta_{k,k+1}$

Montrons le résultat suivant :

C14 : $\forall k < L$

- $c_k = a_k$
- $b_k(m_k - 1) + m_k\, d_k(N_{k+1} - n_{k+1} - 1) = m_k\, a_k^2\, (N_{k+1} - n_{k+1}) - a_k$

Dem : soient $k < L$, $i \in U_{k+1}$, $j \in W_i \cap U_k$

. De C13 et H11, nous avons :

$$N_k - n_k = \beta_{k,k+1} \cdot M_{k,k+1} (N_{k+1} - n_{k+1})$$

$$\Rightarrow \quad n_k = m_k [n_{k+1} + a_k (N_{k+1} - n_{k+1})]$$

. $$\alpha_j = a_k + \alpha_i(1 - a_k) \qquad (c12)$$

. $$\sum_{\substack{j' \in U_k \\ j' \neq j}} \alpha_{jj'} = \alpha_j(n_k - 1) \qquad (Th1)$$

$$= a_k(m_k\, n_{k+1} - 1) + m_k\, a_k^2(N_{k+1} - n_{k+1})$$
$$+ \alpha_i[m_k\, n_{k+1} - 1 + a_k + m_k\, a_k(N_{k+1} - 2\, n_{k+1})$$
$$- m_k\, a_k^2(N_{k+1} - n_{k+1})]$$

. Si $m_k \geq 2$:

$$\sum_{\substack{j' \in W_i \cap U_k \\ j' \neq j}} \alpha_{jj'} = \sum_{\substack{j' \in W_i \cap U_k \\ j' \neq j}} \left(\alpha_i + b_k(1-\alpha_i)\right) \qquad (C8\text{-}H12)$$

$$= (m_k - 1)(\alpha_i + b_k(1-\alpha_i)) \qquad (H4)$$

qui est valable aussi dans le cas où $m_k = 1$ puisqu'alors le rameau i ne porte qu'un seul rameau : j et :

$$\sum_{\substack{j' \in W_i \cap U_k \\ j' \neq j}} \alpha_{jj'} = 0$$

. $$\sum_{\substack{i' \in U_{k+1} \\ i' \neq i}} \sum_{j' \in W_{i'} \cap U_k} \alpha_{jj'}$$

$$= m_k \sum_{\substack{i' \in U_{k+1} \\ i' \neq i}} [\alpha_{ii'} + (\alpha_i - \alpha_{ii'})(c_k + (\alpha_{i'} - \alpha_{ii'})\, c_k +$$
$$(1 - \alpha_i - \alpha_{i'} + \alpha_{ii'})\, d_k] \qquad (H4\text{-}C10)$$

$$= m_k\, d_k(N_{k+1} - n_{k+1} - 1) + m_k\, n_{k+1}\, c_k$$
$$+ m_k\, \alpha_i[n_{k+1} - 1 + c_k(N_{k+1} - 2\, n_{k+1}) - d_k(N_{k+1} - n_{k+1} - 1)]$$

. D'autre part, de C6, nous avons :

$$\sum_{\substack{j'\in U_k \\ j'\neq j}} \alpha_{jj'} = \sum_{i\in U_{k+1}} \sum_{\substack{j'\in W_i,\cap U_k \\ j'\neq j}} \alpha_{jj'}$$

$$= \sum_{\substack{j'\in W_i\cap U_k \\ j'\neq j}} \alpha_{jj'} + \sum_{\substack{i'\in U_{k+1} \\ i'\neq i}} \sum_{j'\in W_i,\cap U_k} \alpha_{jj'}$$

D'où, pour tout i de U_{k+1} :

$$(1-\alpha_i)[m_k a_k^2 (N_{k+1}-n_{k+1}) - b_k(m_k-1) - m_k d_k (N_{k+1}-n_{k+1}-1)$$
$$- a_k + (c_k - a_k) m_k n_{k+1}] = \alpha_i m_k (a_k - c_k)(N_{k+1}-n_{k+1}).$$

D'où le résultat.

R10 : Si les plans de sondages sont à tailles aléatoires ν_k , en suivant le même raisonnement, on aboutit à :

$$\alpha_i m_k c_k N_{k+1} + (1-\alpha_i)[a_k + b_k (m_k - 1) + m_k d_k (N_{k+1} - 1)]$$
$$= E_{p_k} 1_j(S_k)\nu_k - m_k E_{p_{k+1}} \{(c_k - d_k)\nu_{k+1} - (1 - 2c_k + d_k) 1_i(S_{k+1})\nu_{k+1}\}$$

Nous ne pouvons plus appliquer C13. En sommant sur les j de U_k , nous avons :

$$E_{p_{k+1}} \{m_k(N_{k+1} - \nu_{k+1})\ a_k + b_k(m_k - 1) + m_k\ d_k(N_{k+1} - \nu_{k+1} - 1) + 2\ m_k c_k \nu_{k+1}] \}$$
$$= E_{p_k} \nu_k^2 - m_k E_{p_{k+1}} \nu_{k+1}^2$$

Je n'ai rien pu en déduire.

R11 : $\forall k < L$, $\forall i\neq i' \in U_{k+1}$, $\forall j \in W_i \cap U_k$, $\forall j' \in W_i, \cap U_k$

$$c_k = a_k \iff P(j \in S_k\ /_i \notin S_{k+1}) = P(jj' \in S_k\ /_i \notin S_{k+1}, i' \in S_{k+1})$$

Donc : $P(j \in S_k\ /_i \notin S_{k+1}, i' \in S_{k+1}) = P(j \in S_k\ /_i \notin S_{k+1})$

C15 :

i) $a_k \geq Sup(b_k , d_k)$

$n_k \geq m_k\, n_{k+1} + 1$

ii) . $\forall i \in U_{k+1}$, $\forall j \neq j' \in W_i \cap U_k$

$p(jj' \notin S_k / i \notin S_{k+1}) = 1 - 2\, a_k + b_k$

. $\forall i \neq i' \in U_{k+1}$, $\forall j \in W_i \cap U_k$, $\forall j' \in W_{i'} \cap U_k$

$P(jj' \notin S_k / ii' \notin S_{k+1}) = 1 - 2\, a_k + d_k$

iii) Si $b_k = 0$ alors

$m_k\, n_{k+1} + 1 \leq n_k \leq N_{k+1} + (m_k - 1)\, n_{k+1}$

Dem : . le fait que $a_k \geq b_k$ découle de leurs définitions, alors que $a_k \geq d_k$ découle de H9.

De C14, nous avons :

$a_k + b_k(m_k - 1) + m_k\, d_k(N_{k+1} - n_{k+1} - 1) = m_k\, a_k^2(N_{k+1} - n_{k+1})$

$b_k \geq 0$ et $d_k \geq 0 \Rightarrow$

$m_k\, a_k^2\, (N_{k+1} - n_{k+1}) - a_k \geq 0$

d'où i) en appliquant C13.

. $\forall i \in U_{k+1}$, $\forall j \neq j' \in W_i \cap U_k$

$P(j \in S_k , j' \notin S_k / i \notin S_{k+1})$

$= P(j \in S_k / i \notin S_{k+1}) - P(jj' \in S_k / i \notin S_{k+1})$

$= a_k - b_k$

$P(jj' \notin S_k / i \notin S_{k+1})$

$= 1 - P(j \in S_k , j' \notin S_k / i \notin S_{k+1})$

$- P(j \notin S_k , j' \in S_k / i \notin S_{k+1})$

$- P(jj' \in S_k / i \notin S_{k+1})$

$= 1 - 2\,(a_k - b_k) - b_k = 1 - 2\, a_k + b_k$

la deuxième égalité se démontre de la même façon en appliquant R11.

. Si $b_k = 0$, de C14 nous avons :

$$d_k = a_k - a_k \frac{1 - m_k a_k}{m_k (N_{k+1} - n_{k+1} - 1)}$$

$$d_k \leq a_k \Rightarrow a_k \leq \frac{1}{m_k}$$

d'où iii) en appliquant C13.

R12 : En déterminant les coefficients a_k , b_k , d_k vérifiant C14 et C15, les probabilités d'inclusion d'ordre 1 et 2 de U_k sont complétement définies en fonction de celles de la population U_{k+1} .

Le coefficient a_k est lié directement à la taille de l'échantillon (C13), tandis que b_k et d_k en contrôlent la "dispersion". Un cas particulier est $b_k = 0$: on ne prélève pas deux rameaux d'âge k portés par le même élément de U_{k+1} sauf si ce dernier l'a été.

C'est un cas intéressant étant donné la contrainte d'inclusion H9 qui nous rapporche d'un échantillonnage par grappes. Nous verrons à la fonction coût l'intérêt d'un tel cas, qui permet de limiter les dommages causés au niveau d'un arbre.

b) Nous poursuivons le même but, mais ici pour les 2 populations U_h *et* U_ℓ, *pour tout* $h \leq \ell$

A partir de H9 et C15, nous avons :

$$\underline{C16} : \begin{cases} \forall h \leq \ell \ , \forall i \neq i' \in U_\ell \ , \forall j \in W_i \cap U_h \ , \forall j' \in W_i \cap U_h \\ P(jj' \notin S_h \ / \ ii' \notin S) = \prod_{k=h}^{\ell-1} (1 - 2\,a_k + d_k) \end{cases}$$

$\underline{Dem}$: $\forall k : h \leq k \leq \ell$

$\forall r \in W_i \cap U_k$ *tel que* $j \in W_r$

$\forall r' \in W_i \cap U_k$ *tel que* $j' \in W_{r'}$

$P(jj' \notin S_h) = P(jj' \notin S_h \cap r\, r' \notin S_k)$ *(H9)*

Soit :

$P(jj' \notin S_h \ / \ ii' \notin S_\ell) = P(jj' \notin S_h \ / \ rr' \notin S_k) \cdot P(rr' \notin S_k \ / \ ii' \notin S_\ell)$

Il suffit alors d'appliquer C15-ii de proche en proche

Notons alors :

. $\forall h \leq k \ , \forall i \neq i' \in U_k \ , \forall j \in W_i \cap U_h \ , \forall j' \in W_{i'} \cap U_h$

$\gamma_{hk} = P(jj' \notin S_h \ / \ ii' \notin S_k)$

$\gamma_h = \gamma_{hL}$

. $\forall k < L \ , \forall i \in U_{k+1} \ , \forall j \neq j' \in W_i \cap U_k$

$\delta_k = P(jj' \notin S_k \ / \ i \notin S_{k+1})$

De C14 et C15, nous avons :

$\underline{C17}$:

. $\forall k < L$

$Sup\ (0,\ 2\,\beta_{k,k+1} - 1) \leq \delta_k \leq \beta_{k,k+1}$

$$\beta_{k,k+1}\ \frac{\beta_{k,k+1}(N_{k+1} - n_{k+1}) - 1}{N_{k+1} - n_{k+1} - 1} \leq \gamma_{k,k+1} \leq \beta_{k,k+1}$$

$$\delta_k\,(m_k - 1) + m_k\,\gamma_{k,k+1}\,(N_{k+1} - n_{k+1} - 1)$$

$$= \beta_{k,k+1}\,[m_k\,\beta_{k,k+1}\,(N_{k+1} - n_{k+1}) - 1]$$

. $\forall h \leq k \leq \ell \qquad \gamma_{h\ell} = \gamma_{hk} \cdot \gamma_{k\ell}$

. $1 = \gamma_L \geq \gamma_{L-1} \geq \ldots \geq \gamma_1 \geq 0$

On en déduit, pour tout $h \leq \ell$, les probabilités d'inclusion d'ordre 2 de U_h en fonction de celles de U_ℓ :

<u>C18</u> :

i) . $\forall h \leq \ell$, $\forall i \neq i' \in U_\ell$, $\forall j \in w_i \cap U_h$, $\forall j' \in w_{i'} \cap U_h$

$$P(jj' \notin S_h) = \gamma_{h\ell}\, P(ii' \notin S_\ell)$$

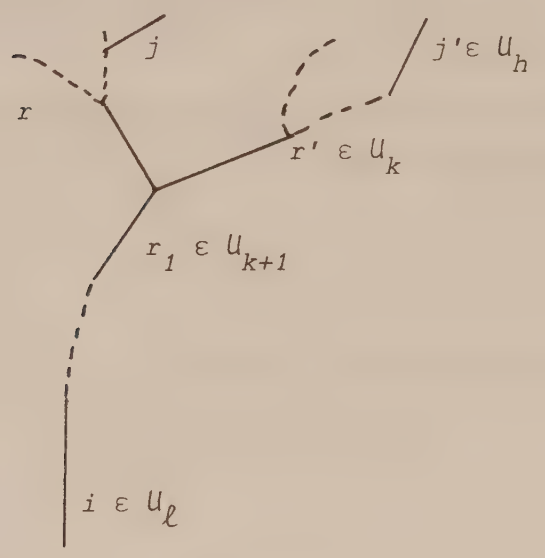
ii) . $\forall h \leq k < \ell$, $\forall i \in U_\ell$, $\forall r_1 i \in w_i \cap U_{k+1}$, $\forall r \neq r' \in w_{r_1} \cap U_k$

$\forall j \in w_r \cap U_h$, $\forall j' \in w_{i'} \cap U_h$

$$P(jj' \notin S_h) = \gamma_{hk}\, \delta_k\, \beta_{k+1,\ell}\, P(i \notin S_\ell)$$

i) correspond au cas de deux rameaux d'âge h portés par 2 rameaux différents d'âge ℓ

ii) est le cas où j et j' sont portés par le même rameau i d'âge ℓ, k étant l'âge à partir duquel ils sont portés par des rameaux différents.



<u>Dem</u> : *i) découle de la définition de* γ_{hk}

ii) de i) nous avons :

$$P(jj' \notin S_h) = \gamma_{hk}\, P(rr' \notin S_k)$$

$$C15 - ii \Rightarrow P(rr' \notin S_k) = \delta_k\, P(r_1 \notin S_{k+1})$$

$$C12 \Rightarrow P(r_1 \notin S_{k+1}) = \beta_{k+1,\ell}\, P(i \notin S_\ell)$$

De C12 et C18, nous avons, pour les mêmes cas que C18 :

$$\underline{C19} : \begin{cases} i)\ \forall h \leq \ell\ ,\ \forall i \neq i' \in U_\ell\ ,\ \forall j \in W_i \cap U_h\ ,\ \forall j' \in W_{i'} \cap U_h \\ \quad \alpha_{jj'} = 1 - \beta_{h\ell}(1-\alpha_i) - \beta_{h\ell}(1-\alpha_{i'}) + \gamma_{h\ell}(1-\alpha_i-\alpha_{i'}+\alpha_{ii'}) \\ ii)\ \forall h \leq k < \ell\ ,\ \forall i \in U_\ell\ ,\ \forall r_1 \in W_i \cap U_{k+1}\ ,\ \forall r \neq r' \in W_{r_1} \cap U_k \\ \quad \forall j \in W_r \cap U_h\ ,\ \forall j' \in W_{r'} \cap U_h \\ \quad \alpha_{jj'} = 1 - (1-\alpha_i)\ [2\beta_{h\ell} - \gamma_{hk}\ \delta_k\ \beta_{k+1,\ell}] \end{cases}$$

D'autre part, un résultat équivalent à C14, dont la démonstration suit la même démarche est :

$$\underline{C20} : \begin{cases} \forall h < \ell \\ \sum_{k=h\ldots\ell-1} (m_k - 1)\ \dfrac{\delta_k\ \beta_{k+1}}{M_k\ \gamma_k} \\ = \dfrac{\beta_h}{M_h\ \gamma_h}\ [N_h - n_h - 1]\ - \dfrac{\beta_\ell}{M_\ell\ \gamma_\ell}\ [N_\ell - n_\ell - 1] \end{cases}$$

La seule différence avec C14 surgit quand il s'agit de déterminer, pour tout j de U_h porté par i de U_ℓ :

$$\sum_{\substack{j' \in W_i \cap U_h \\ j' \neq j}} \alpha_{jj'}$$

En effet, considérons les rameaux j_k , k=h...ℓ tels que :

$$j_\ell \in U_\ell \quad \text{et} \quad \forall k\ ,\ h \leq k < \ell \quad j_k \in W_{j_{k+1}} \cap U_k$$

alors :

$$\sum_{\substack{j \in W_{j_\ell} \cap U_h \\ j \neq j_h}} \alpha_{jj_h}$$

$$= \sum_{k=h\ldots\ell-1} \sum_{\substack{i \in W_{j_{k+1}} \cap U_k \\ i \neq j_k}} \sum_{j \in W_i \cap U_h} \alpha_{jj_h}$$

$$= \sum_{k=h\ldots\ell-1} M_{hk}(m_k - 1)[1 - (1-\alpha_{j_\ell})(2\beta_{h\ell} - \gamma_{hk}\ \delta_k\ \beta_{k+1,\ell})]$$

En continuant la même démarche, nous arrivons à :

$$\sum_{k=h\ldots\ell-1} M_{hk}(m_k - 1)\,\gamma_{hk}\,\delta_k\,\beta_{k+1,\ell}$$

$$= M_{h\ell}(N_\ell - n_\ell - 1)\,(\beta^2_{h\ell} - \gamma_{h\ell}) - \beta_{h\ell}(1 - M_{h\ell}\beta_{h\ell})$$

Il suffit alors d'appliquer C11, C12, C13 et C17 :

$$\cdot \sum_{k=h\ldots\ell-1} M_{hk}(m_k-1)\,\gamma_{hk}\,\delta_k\,\beta_{k+1,\ell}$$

$$= \frac{M_h\,\gamma_h}{\beta_\ell} \sum_{k=h\ldots\ell-1} (m_k - 1)\,\frac{\delta_k\,\beta_{k+1}}{M_k\,\gamma_k}$$

$$\cdot \frac{\beta_\ell}{M_h\,\gamma_h}\,[M_{h\ell}(N_\ell - n_\ell - 1)(\beta^2_{h\ell} - \gamma_{h\ell}) - \beta_{h\ell}(1 - M_{h\ell}\,\beta_{h\ell})]$$

$$= \frac{N_\ell - n_\ell - 1}{M_\ell}\left(\frac{\beta_h^2}{\beta_\ell\,\gamma_h} - \frac{\beta_\ell}{\gamma_\ell}\right) - \frac{\beta_h}{M_h\,\gamma_h} + \frac{\beta_h^2}{M_\ell\,\beta_\ell\,\gamma_h}$$

$$= \frac{\beta_h}{M_h\,\gamma_h}\left[(N_\ell - n_\ell)\,\frac{\beta_h}{\beta_\ell}\,\frac{M_h}{M_\ell} - 1\right] - \frac{\beta_\ell}{M_\ell\,\gamma_\ell}\,(N_\ell - n_\ell - 1)$$

$$= \frac{\beta_h}{M_h\,\gamma_h}\,[N_h - n_h - 1] - \frac{\beta_\ell}{M_\ell\,\gamma_\ell}\,(N_\ell - n_\ell - 1)$$

c) Il nous reste à exprimer les probabilités d'inclusion α_{ij} de toute paire d'éléments $\{i,j\}$ de la population totale U, en fonction de celle de U_L.

C'est surtout un travail de synthèse de ce qui précède, à part le résultat C21.

Comme $\alpha_{ij} = \alpha_{ji}$, et puisque à tout rameau i de U , nous savons associer un âge k, nous nous intéresserons aux couples (i,j) tels que l'âge h de j soit inférieur ou égal à l'âge k de i.

Dans le cas où $h = k$, C5 résume les différentes possibilités, et C19 donne les probabilités d'inclusion correspondantes (avec $\ell = L$).

Si $h < k$, nous avons 2 cas :

- soit j est porté par i, alors la contrainte d'inclusion H9 implique que :

$$\alpha_{ij} = \alpha_i$$

- soit j n'est pas porté par i, alors C21 donne α_{ij} pour les 2 cas possibles correspondants à ceux de C5 :

  Le cas i) correspond à celui où j et i sont portés par des arbres différents r et r'

  Le cas ii) correspond à celui où j et i sont portés par le même arbre r, ℓ est l'âge à partir duquel ils ne sont plus portés par le même rameau.

  Pour garder les mêmes notations que dans ce qui précède, nous chercherons α_{ji}, à la place de α_{ji}

C21 :

i) $\forall h \leq k$, $\forall r \neq r' \in U_L$, $\forall j \in W_r \cap U_h$, $\forall i' \in W_{r'} \cap U_k$

$$\alpha_{ji'} = 1 - (1 - \alpha_r)\,\beta_h - (1 - \alpha_{r'})\,\beta_k + (1 - \alpha_r - \alpha_{r'} + \alpha_{rr'})\,\beta_{hk}\,\gamma_k$$

ii) $\forall h \leq k < \ell < L \qquad \forall r \in U_L$

$\forall r_1 \in W_r \cap U_{\ell+1}$, $\forall r_2 \neq r_3 \in W_{r_1} \cap U_\ell$

$\forall j \in W_{r_2} \cap U_h$, $\forall i' \in W_{r_3} \cap U_k$

$$\alpha_{ji'} = 1 - (1 - \alpha_r)[\beta_h + \beta_k - \beta_{hk}\,\gamma_{k\ell}\,\delta_\ell\,\beta_{\ell+1}]$$

Dem :

. Montrons d'abord que :

$\forall i \neq i' \in U_k$, $\forall j \in W_i \cap U_h$, $\forall j' \in W_{i'} \cap U_h$

$$P(jj' \in S_h \,/\, i \notin S_k \,,\, i' \in S_k) = 1 - \beta_{hk}$$

Ceci est vrai pour $k = h + 1$ (C14)

Supposons que c'est vérifié à l'ordre $k > h$, et soient $r \neq r' \in U_{k+1}$ tels que $i \in W_r$ et $i' \in W_{r'}$,

Alors :

$$P(jj' \in S_h \,/\, r \notin S_{k+1} \,,\, r' \in S_{k+1})$$

$$= P(jj' \in S_h \cap ii' \in S_k \,/\, r \notin S_{k+1} \,,\, r' \in S_{k+1})$$

$$+ P(jj' \in S_h \cap i \notin S_k \cap i' \in S_k \,/\, r \notin S_{k+1} \,,\, r' \in S_{k+1})$$

$$= P(ii' \in S_k \,/\, r \notin S_{k+1} \,,\, r' \in S_{k+1})$$

$$+ P(jj' \in S_h \,/\, i \notin S_k \,,\, i' \in S_k) \cdot P(i \notin S_k \,,\, i' \in S_k \,/\, r \notin S_{k+1} \,,\, r' \in S_{k+1})$$

(H9)

$$= 1 - \beta_{k,k+1} + (1 - \beta_{hk})\,\beta_{k,k+1} = 1 - \beta_{h,k+1}$$

. Appliquons C10, nous avons alors :

$$P(j \in S_h \cap i' \in S_k) = (\alpha_{i'} - \alpha_{ii'})(1 - \beta_{hk}) + \alpha_{ii'}$$

avec i le rameau porteur de j d'âge k ($i \neq i'$).

. Il suffit alors d'appliquer C12 et C19 pour les 2 cas correspondants.

Si nous supposons que la taille n_L de l'échantillon prélevé de la population U_L (des arbres) est telle que nous pouvons négliger $\frac{1}{N_L - n_L}$, alors nous pouvons résumer les contraintes vérifiées par les paramètres par :

C22 :

- $1 = \beta_L = \gamma_L = \delta_L$
- $\{\beta_k\}_{k=1...L}$, $\{\gamma_k\}_{k=1...L}$, $\{\gamma_k/\beta_k\}_{k=1...L}$ et $\{\beta_k^2/\gamma_k\}_{k=1...L}$ sont positives croissantes
- $\forall h < L$

$$\beta_{h+1} - \beta_h \geq \frac{1}{M_h (N_L - n_L)}$$

$$2\beta_h - \beta_{h+1} \leq \delta_h \cdot \beta_{h+1} \leq \beta_h$$

$$\sum_{k=h...L-1} (m_k - 1) \frac{\delta_k}{M_k} \frac{\beta_{k+1}}{\gamma_k} = 1 - \frac{\beta_h}{M_h \gamma_h} - (N_L - n_L)\left[1 - \frac{\beta_h^2}{\gamma_h}\right]$$

la dernière égalité découle directement de C20.

Les contraintes sur les $\{\beta_k\}$ sont équivalentes à celles de C11 et de C15-i. Celles sur les $\{\delta_k\}$ à C14 ou encore à C17.

D'autre part, $\frac{\gamma_h}{\beta_h} \leq \frac{\gamma_{h+1}}{\beta_{h+1}} \iff \gamma_{h,h+1} \leq \beta_{h,h+1}$ *(C17)*

il nous reste à montrer que si $\frac{1}{N_L - n_L}$ *est négligeable alors :*

$$\gamma_{h,h+1} \geq \beta_{h,h+1} \frac{\beta_{h,h+1} (N_{h+1} - n_{h+1}) - 1}{N_{h+1} - n_{h+1} - 1} \qquad (C17)$$

est équivalente à : $\frac{\beta_h^2}{\gamma_h} \leq \frac{\beta_{h+1}^2}{\gamma_{h+1}}$

Pour ce, il suffit d'appliquer C13 à l'expression de C17. Nous avons alors :

$$\gamma_{h,h+1} \geq \beta_{h,h+1} \frac{M_{h+1} (N_L - n_L) \beta_h - 1}{M_{h+1} (N_L - n_L) \beta_{h+1} - 1}$$

soit $\gamma_{h,h+1} \geq \beta_{h,h+1}^2$ *en négligeant* $\frac{1}{M_h (N_L - n_L)}$ *d'où le résultat.*

Cas particuliers :

i) Si nous nous imposons de ne pas prélever deux rameaux de même âge k(k ⩾ h), portés par le même rameau d'âge k+1, à moins que ce dernier n'ait été prélevé, ceci équivaut à poser que :

$$\forall k : h \leqslant k < L \qquad \delta_k = \beta_k / \beta_{k+1}$$

De C17, nous avons alors :

$$\gamma_{k,k+1} = \beta^2_{k,k+1} - \beta_{k,k+1} \frac{1 - \beta_{k,k+1}}{N_{k+1} - n_{k+1} - 1}$$

Soit :

$$\frac{\beta_k}{\gamma_k} [\beta_k M_{k+1} (N_L - n_L) - 1]$$

$$= \frac{\beta_{k+1}}{\gamma_{k+1}} [\beta_{k+1} M_{k+1} (N_L - n_L) - 1]$$

En remplaçant dans C22, et en tenant compte de C15, iii, nous avons :

C23 :

Si $\forall h < L \qquad \delta_h = \beta_h / \beta_{h+1}$

alors les paramètres β_h *vérifient :*

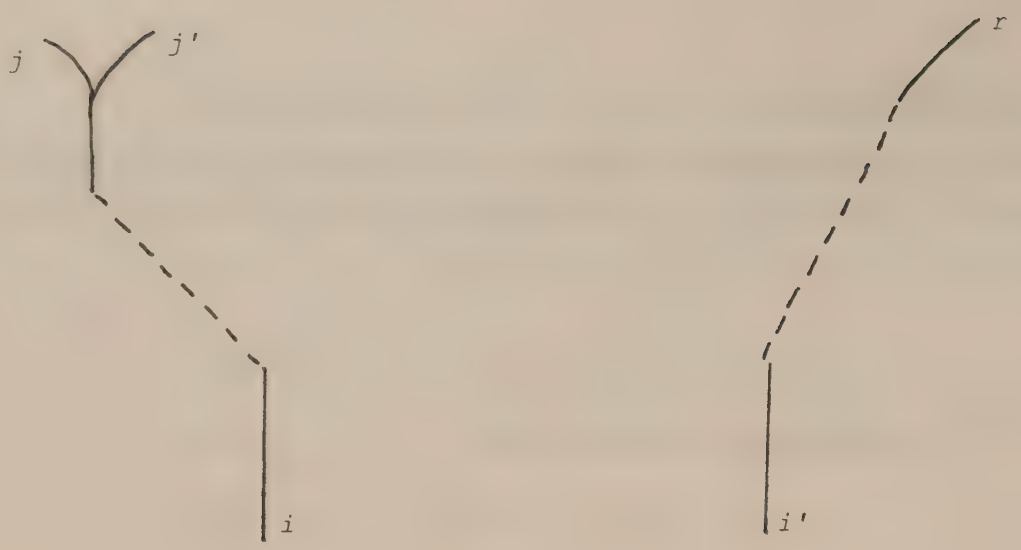
$$\frac{1}{M_h (N_L - n_L)} \leqslant \beta_{h+1} - \beta_h \leqslant \frac{1}{m_h} \beta_{h+1}$$

et les paramètres γ_h *sont entièrement déterminés par :*

$$1 - \frac{\beta^2_h}{\gamma_h} = \sum_{k=h+1 \ldots L} \frac{\beta^2_k}{\gamma_k} \frac{\beta_{k-1} - \beta_k}{\beta_{k-1} M_k (N_L - n_L) - 1}$$

Ce cas est très intéressant. En dispersant l'échantillon prélevé parmi les rameaux d'âge k, sur le maximum de rameaux porteurs d'âge k+1, nous limitons les dommages causés au niveau d'un arbre.

ii) : soit k un âge quelconque, et soient j, j' et r des rameaux de U_k *tels que j et j' soient portés par le même rameau de* U_{k+1} *, donc par le même arbre i, alors que r est porté par un arbre i' distinct de i.*



Alors : $\frac{\delta_k \beta_{k+1}}{\gamma_k} = \frac{P(jj' \notin s_k / i \notin s_L)}{P(j\,r \notin s_k / ii' \notin s_L)}$

de C22, nous avons :

$$\frac{2\beta_k - \beta_{k+1}}{\gamma_k} \leq \frac{\delta_k \beta_{k+1}}{\gamma_k} \leq \frac{\beta_k}{\gamma_k}$$

Supposer ce rapport constant quand l'âge varie, revient à avoir la même attitude vis-à-vis de la dispersion de l'échantillon sur les arbres (et non, comme dans le cas i, vis-à-vis des rameaux porteurs). Nous avons alors :

$$k = 1 \ldots L-1 \quad \frac{\delta_k \beta_{k+1}}{\gamma_k} = \frac{\delta_{L-1}}{\gamma_{L-1}}$$

et de C22, nous déduisons :

$$\gamma_k = \gamma_{L-1}(M_{L-1} - 1) \frac{\beta_k [M_k \beta_k (N_L - n_L) - 1]}{\delta_{L-1}(M_k - 1) + \gamma_{L-1} M_k (N_L - n_L - 1)}$$

Les paramètres $\{\beta_k, \gamma_k, \delta_k\}_{k=1 \ldots L-1}$ sont alors entièrement déterminés par la donnée des $(\beta_k)_{k=1 \ldots L-1}$ et de γ_{L-1} avec :

$$\beta_{L-1} \frac{\beta_{L-1}(N_L - n_L) - 1}{N_L - n_L - 1} \leq \gamma_{L-1} \leq \beta_{L-1}$$

B - II - 5) Liens entre plans de sondage et probabilités d'inclusion

Nous dirons qu'un plan de sondage P (abréviation de (S,a,P)) est complétement déterminé si tous les échantillons possibles s de S ainsi que leurs probabilités P(S=s) sont donnés avec :

$$\forall s \in S \qquad 0 \leq P(S=s) \leq 1$$

$$\text{et} \qquad \sum_{s \in S} P(S=s) = 1$$

Dans notre cas, P vérifie les hypothèses H6, H7, H8 et H11. Il est alors complétement déterminé si toutes les probabilités d'inclusion $\alpha_{i_1 \ldots i_n}$ le sont avec :

$$n = \sum_{k=1 \ldots L} n_k = \text{taille de } P \qquad (H11)$$

et $i_1 \ldots i_n$ = n éléments distincts de $\mathcal{U}$ comprenant n_k rameaux d'âge k, pour k = 1...L

donc aussi toutes les probabilités d'inclusion d'ordre m, m = 1 ... n-1 pour tous les éléments $i_1 \ldots i_m$ tous distincts et comprenant au plus n_k rameaux d'âge k, k = 1 ... L

Déterminer complétement P revient donc à se donner les :

$$\left(\prod_{k=1 \ldots L} C_{N_k}^{n_k} \right) \text{ couples } \left(s \ , \ P(S=s)\right)$$

Dans un premier temps,
Une démarche moins laborieuse consiste à déterminer plutôt un schéma d'échantillonnage engendrant P . (Th2).

Nous définissons un schéma d'échantillonnage $\{\Pi_i(j)\}_{\substack{i=1 \ldots n \\ j=1 \ldots N}}$ comme un processus de sélection des individus un à un de la population totale $\mathcal{U}$, vérifiant :

$$\forall i = 1 \ldots n \qquad \forall j = 1 \ldots N \qquad 0 \leq \Pi_i(j) \leq 1$$

$$\text{et} \quad \sum_{j=1 \ldots N} \Pi_i(j) = 1$$

et fonctionnant comme suit :

- tirer un 1er individu avec la probabilité $\Pi_1(j_1)$

- Pour $k = 2 \ldots n$, tirer l'individu j_k , connaissant les résultats $j_1 \ldots j_{k-1}$ des prélèvements 1 ... k-1 avec la probabilité :

$$\frac{\Pi_k(j_k)}{1 - \Pi_1(j_1) - \ldots - \Pi_{k-1}(j_{k-1})}$$

Dans notre cas $j_1 \ldots j_k$ sont tous distincts (H7)

La probabilité de prélever la séquence ordonnée $(j_1 \ldots j_k)$, pour $k = 1 \ldots n$, est :

$$\tau(j_1 \ldots j_k) = \Pi_1(j_1) \frac{\Pi_2(j_2)}{1 - \Pi_1(j_1)} \ldots \frac{\Pi_k(j_k)}{1 - \Pi_1(j_1) - \ldots - \Pi_{k-1}(j_{k-1})}$$

et la probabilité de prélever l'échantillon $\{j_1 \ldots j_n\}$ (non ordonné : H6) est :

$$\alpha_{j_1 \ldots j_n} = P(S = \{j_1 \ldots j_n\}) = \sum_{\phi} \tau(j_{\phi(1)} \ldots j_{\phi(n)})$$

où la somme est effectuée sur toutes les permutations ϕ de $\{1 \ldots n\}$

Nous avons :

$$\forall i \in U \qquad \alpha_i = \sum_{k=1\ldots n} \sum_{(k-1,i)} \tau(j_1, \ldots, j_{k-1}, i)$$

et $\forall i \neq i' \in U$

$$\alpha_{ii'} = \sum_{k=2\ldots n} \sum_{\ell=1\ldots k-1} \sum_{(k-2,\ell,i,i')} \tau(j_1,\ldots,j_{\ell-1},i',j_{\ell+1},\ldots,i)$$

$$+ \sum_{k=1\ldots n-1} \sum_{\ell=k+1\ldots n} \sum_{(\ell-2,k,i',i)} \tau(j_1,\ldots,j_{k-1},i,j_{k+1},\ldots,i')$$

avec $\sum_{(k-1,i)}$ désignant la somme sur tous $(j_1,\ldots,j_{k-1})$ tous distincts et tous différents de i

et $\sum_{(k-2,\ell,i,i')}$ la somme sur tous les $(j,\ldots,j_{\ell-1}, j_{\ell+1}, \ldots j_{k-1})$ tous distincts et tous différents de i et i'

T.V.H. RAO a montré le Th suivant, valable pour tout plan de sondage pourvu que tous les individus de la population soient répertoriés :

Th2 : On a une correspondance biunivoque entre les plans de sondage complétement déterminés et les schémas d'échantillonnage.

Dans notre cas, on s'est fixé seulement les probabilités d'inclusion d'ordre 1 et 2, puisque c'est les seules qui interviennent lors du calcul des espérances et variances de toutes les fonctions linéaires de S.

Nous avons donc plusieurs possibilités pour compléter la détermination du plan de sondage ou plutôt du schéma d'échantillonnage tout en respectant H6, H7, H8 et H9.

Un tel choix sera fait en fonction d'autres critères, faisant intervenir l'estimateur choisi. Nous l'illustrons néanmoins par l'exemple simple où toutes les probabilités d'inclusions d'ordre 1 sont fixées :

$$\forall\ i \in U \qquad \alpha_i = \frac{n}{N}$$

où n est la taille fixe de l'échantillon sans remise prélevée. La solution la plus utilisée est :

- *le plan randomisé simple :*

$$\forall\ s \in S \qquad P(S=s) = \begin{cases} 0 & si\ \ \nu(s) \neq n \\ \dfrac{n!(N-n)!}{N!} & si\ \ \nu(s) = n \end{cases}$$

le schéma d'échantillonnage est déterminé par :

$$\Pi_i(j) = \frac{n - i + 1}{N - i + 1} \quad pour \quad i = 1 \ldots n\ ,\ j = 1 \ldots N$$

et les probabilités d'inclusion de 1er et 2ème ordre sont :

$$\forall\ i \in U \qquad \alpha_i = \frac{n}{N}$$

$$\forall\ i \neq j \in U \qquad \alpha_{ij} = \frac{n(n-1)}{N(N-1)}$$

D'autres schémas d'échantillonnage, non basés sur les tirages un à un des individus, ont été définis. C'est le cas pour l'échantillonnage systèmatique, stratifié, ou encore le cas du plan du sondage défini par RAO - HARTLEY - COCHRAN (1962). Le Th2 s'énonce alors :

> *Pour tout plan de sondage, il existe au moins un schéma d'échantillonnage qui l'engendre.*

Une deuxième solution du cas précédent, qui est aussi très utilisée est :

- l'échantillonnage systématique (taille fixe n)

Nous supposons pour simplifier que $m = N/n$ est un entier.

U est subdivisée et rangée arbitrairement en n sous populations de taille m chacune. Donc pour tout individu i de U, on connait la sous population à laquelle il appartient, et son numéro de rangement $\Theta(i)$ dans cette sous population.

Un nombre j, $1 \leq j \leq m$, est déterminé au hasard. Alors :

$\forall i \in U$ si $\Theta(i) = j$, i est prélevé

sinon , i n'est pas prélevé

Nous avons bien :

$$\forall i \in U \quad \alpha_i = \frac{1}{m} = \frac{n}{N}$$

Par contre, nous avons :

$$\forall i \neq j \in U \quad \alpha_{ij} = \begin{cases} 0 & si \quad \Theta(i) \neq \Theta(j) \\ 1 & sinon \end{cases}$$

et

$$\forall s \in S \quad P(S=s) = \begin{cases} \frac{1}{m} = \frac{n}{N} & si \quad \nu(s) = n \\ 0 & si \quad \nu(s) \neq n \end{cases}$$

Pour la simulation, nous avons considéré le cas :

$$\forall i \in U_L \quad \alpha_i = \frac{n_L}{N_L}$$

$$\forall k = 1...L-1 \quad \delta_k = \beta_k / \beta_{k+1} \quad \text{(dispersion maximale)}$$

pour différentes valeurs des β_k

Le schéma d'échantillonnage considéré peut être résumé par l'organigramme :

- Etape 0 : $k = L$

$U'_k = U_k \quad n_{k+1} = 0 \quad so = \emptyset$

- *Etape 1 : prélever un échantillon* s_k *de taille* $(n_k - m_k \cdot n_{k+1})$ *de la population* U'_k *selon un plan "dispersé" donné (voir C23).*

- *Etape 2 :* $s_k = s_k \cup s_o$

  si $k = 1$ *fin*

- *Etape 3 : mettre dans* s_o *tous les individus d'âge (k-1) qui sont portés par les éléments de* s_k

- *Etape 4 :* $U'_{k-1} = U_{k-1} - s_o$

  . $k = k - 1$

  . *retour à l'étape 1*

  Pour plus de précisions, voir la simulation.

B I B L I O G R A P H I E

[1] C. ASOK, B.V. SUKHATME (1976)

On Sampford procedure of uniqual probability Sampling without replacement
J.A.S.A. 71, p. 912-918

[2] C.M. CASSEL, C.E. SARNDAL, J.H. WRETMAN (1977)

Foundations of Inference in survey sampling
Wiley - Interscience

[3] W.G. COCHRAN (1962)

Sampling Techniques
John Wiley & Sons

[4] W.A. ERICSON (1965)

Optimum stratified sampling using prior information
J.A.S.A. 60, p. 750-771

[5] T.V.H. RAO (1962)

An existence theorem in sampling Theory
Sankya A24, 327-330

[6] H. SOLOMON, S. ZACKS (1970)

Optimal design of sampling from finite populations : A critical review and indication of new research areas
J.A.S.A. 65, p. 653-677

[7] W.E. WATERS (1955)

Sequentiel sampling in Forest insect surveys
Forest Sciences 1, 68-80

B - III - INTRODUCTION A L'ESTIMATION

Nous fixons des notations en respectant les deux règles suivantes :

- Une variable aléatoire est notée par une lettre majuscule (S, D, Y,...), et une réalisation particulière par la lettre minuscule correspondate (s, d, y,...).

- les notations correspondant aux sous-populations U_k, k=1...L, se déduisent de celles de la population totale U, en rajoutant l'indice "âge" k.
  Tant qu'il n'y a pas de confusion possible, cet indice est omis.

Dans 1), nous présentons le modèle à population fixe et le modèle de superpopulation, ainsi que la démarche, générale.

Pour estimer $Y_m = \sum_{k=1...L} M_k \bar{Y}_k$, nous considérons :

$$T = \sum_{k=1...L} M_k T_k$$

où T_k est un estimateur de $\bar{Y}_k$, k = 1...L.

Dans 2), nous nous intéressons au modèle à population fixe. Des justifications des hypothèses H6, H7 et H10 sont données.

Dans 3), nous faisons des remarques générales sur le modèle de superpopulation.

La fonction coût, commune aux deux modèles, est explicitée dans 4).

B-III-1) Observations générales :

a) Soit $\Omega = \{d \;/\; s \in S \;,\; y \in R^N\}$

avec d, comme défini en II-1,

$$d = \{(i,\ y_i) \;;\; i \in s\}$$

(Nous supposons que l'information auxiliaire x est complétement déterminée).

Ω *est muni de la* σ *- algèbre* a x B *ou* a *est la* σ *- algèbre des toutes les parties de* S, *et* B *la* σ *- algèbre engendrée par les boréliens de* R^N.

Notons ξ *la distribution jointe de* $Y_1 \ldots Y_n$.

(Ω, a x B*) muni de la mesure produit (P,* ξ*) est un espace probabilisé.*

Définissons la variable aléatoire :

$$D = \{(i,\ Y_i) \;;\; i \in S\}$$

Une fonction T définie sur Ω *est un estimateur de F(Y), si T ne dépend de* $Y_1 \ldots Y_N$ *qu'à travers D et nous sert à prédire F(Y), où F est une fonction quelconque Y.*

Le choix d'un estimateur T de Ym (I-2) est fait en fonction du couple (P , ξ*), avec :*

- *P le plan de sondage, vérifiant les hypothèses et conclusion de II, mais pas entièrement déterminé.*

- ξ *la distribution jointe de* $Y_1 \ldots Y_N$ *, fixe mais inconnue.*

Nous pouvons intervenir au niveau du choix de P et T, alors que ξ *est déterminée par la variable étudiée. Donc, nous avons à déterminer le couple (P,T) appelé stratégie. Nous procéderons en deux étapes :*

- *Aucune connaissance a priori sur la distribution* ξ *n'est utilisée. Les valeurs prises par la variable considérée sont supposées fixes mais inconnues. C'est le modèle à population fixe ou modèle classique.*

- Nous formalisons nos connaissances a priori sur la variable considérée par un ensemble de conditions que ξ est supposée satisfaire. Nous dirons que l'on définit ainsi un "modèle de superpopulation" ou tout simplement modèle M. (Si M décrit une famille paramétrique, et qu'un modèle est également défini pour décrire nos connaissances a priori sur ces paramètres, alors nous retrouvons le modèle bayésien).

Dans 4), nous définissons une fonction coût ne dépendant que du plan de sondage $C(P)$, dans 2) et 3) une fonction risque, correspondant à chacun des cas, dépendant de l'écart entre l'estimation T et la vraie valeur à estimer.

Pour choisir entre deux stratégies $H = (P,T)$ et $H' = (P', T')$, nous avons deux possibilités :

- H est meilleure que H' si :

  $\lambda C(P) + R(H) \leqslant \lambda C(P') + R(H')$

  où λ est un paramètre donné

- P et P' sont tels que $C(P)$ et $C(P')$ soient au plus égaux à un coût total donné.

  Alors H est meilleure que H' si :

  $R(H) \leqslant R(H')$

Pour le modèle à population fixe (resp. de superpopulation) l'inégalité est valable pour tout y de R^N (resp. ξ de M) stricte pour au moins un y de R^N (resp. ξ de M).

Nous utiliserons le deuxième critère. Il tient mieux compte des contraintes imposées par les possibilités matérielles dont on dispose, et connues généralement à l'avance.

Le problème est donc de déterminer une stratégie globale $H = (P,T)$ telle que H soit meilleure que toute autre stratégie (H est dite alors optimale). Dans le cas où une telle stratégie n'existe pas, considérer toutes les stratégies H telles qu'il n'existe pas de stratégie H' meilleure que H (H est dite admissible).

b) Un estimateur T est dit linéaire si, pour tout échantillon s de S, il existe des coefficients b_{os} et b_{is}, pour tout i de s, tels que :

$$T(D/_{S=s}) = b_{os} + \sum_{i \in s} b_{is}\, Y_i$$

T est linéaire homogène si en plus : $b_{os} = 0 \quad s \in S$

Dans la suite , nous considérons les liens entre l'estimation de y_m *et celle des* $\bar{y}_k$ *, k = 1...L*

Nous reprenons les notations précédentes en rajoutant l'indice "âge".

Tout estimateur T de Y_m *peut toujours s'écrire :*

$$T(D) = \sum_{k=1...L} M_k\, T_k(D)$$

Pour k = 1...L, $T_k(D)$ *convient à la définition d'un estimateur de* $\bar{Y}_k$ *, puisqu'il ne dépend de* $Y_{k1} \ldots Y_{kN_k}$ *qu'à travers D, et donc qu'à travers* D_k*. Mais* T_k *peut dépendre des* D_ℓ *,* $\ell \neq k$*.*

Supposons que T soit linéaire, alors :

$\forall\, s \in S$, $\exists\, b_{os}$ *et* $\forall\, i \in s$, $\exists\, b_{is}$ *tels que :*

$$t(d) = T(D/Y=y\ ,\ S=s)$$

$$= b_{os} + \sum_{k=1...L} \sum_{i \in s_k} b_{is}\, Y_{ki} \qquad \forall\, y \in R^N$$

Si nous supposons en plus que :

$$\forall\, s \in S\ ,\ (s = \bigcup_{k=1...L} s_k)\ ,\ \forall_{k=1...L},\ \forall\, i \in s_k\ :$$

$$b_{is} = b_{is_k}$$

alors $T(D) = \sum M_k\, T_k(D_k)$

avec $\forall\, s_k \in S_k$

$$t_k(d_k) = T_k(D_k/_{S_k=s_k}\ ,\ Y_k = y_k)$$

$$= b'_{os_k} + \frac{1}{M_k} \sum_{i \in s_k} b_{is_k}\, y_{ki} \qquad \forall\, y_k \in R^{N_k}$$

pour tous les b'_{os_k} *tels que :*

$$\forall\, s \in S \qquad \sum_{k=1:..L} M_k\, b'_{os_k} = b_{os}$$

Dans le modèle de superpopulation, nous ne considérons que les estimateurs T de Y_m de la forme :

$$T = \sum_{k=1...L} M_k T_k$$

Les estimateurs T_k considérés sont d'abord fonction seulement des D_k , puis des D_h , $h < k$. Dans le second cas, D_h , $h < k$, est considérée comme une information auxiliaire.

Il est évident que cette information auxiliaire ne peut intervenir au niveau de la détermination du plan de sondage (II-3), puisque tous les D_h ne sont déterminés qu'après avoir effectué les prélèvements.

D'autre part, les $D_h (h < k)$ ne constituent pas vraiment une information auxiliaire dans le sens classique du terme : elles ne sont pas connues sur la totalité de la population, mais elles apportent des informations sur des rameaux d'âge k qui n'ont pas été prélévés (H9). C'est aussi la raison pour laquelle nous n'avons pas considéré tous les D_h , $h > k$.

B-III - 2) Modèle à population fixe :

a) Dans ce modèle, les $(y_i)_{i=1...N}$ sont considérés fixés mais inconnus. La seule variable aléatoire dans D est l'échantillon S :

$$D = \{(i, y_i) , i \in S\}$$

Pour toute fonction F(y) sur Ω, notons :

$$E_p\, T(D) = \sum_{s \in S} T(d) \,.\, P(S=s)$$

$$V_p\, T(D) = E_p \left(T(D) - E_p\, T(D)\right)^2$$

$$R(P,T) = R(H) = E_p \left(T(D) - F(y)\right)^2$$

T est un estimateur P-sans biais de F(y) si :

$$E_p\, T(D) = F(y) \qquad \forall\, y \in R^N$$

R13 : Si T est un estimateur P-sans biais de F(y) alors $R(P,T)=V_p\, T(D)$

R14 : Dans le cas des sous populations U_k , $k = 1...L$

$$F(y_k) = \bar{y}_k \quad \text{et} \quad R(P_k , T_k) = E_{p_k} (T_k(D_k) - \bar{y}_k)^2$$

Dans le cas de la population totale U

$$F(y) = y_m \quad \text{et} \quad R(P , T) = E_p (T(D) - y_m)^2$$

Deux résultats sur l'estimabilité de $\bar{y}$ (GODAMBE 1955) et de toute fonction quadratique des y_i , $i = 1...N$, donc en particulier de la variance et du risque (HANURAV 1966) sont :

Th3 : Pour tout plan de sondage P, une fonction linéaire des y_i , $i = 1...N$

$$F = \ell_o + \sum_{i=1} \ell_i\, y_i$$

est estimable si et seulement si

$$\alpha_i > 0 \quad \text{pour tout} \quad i \quad \text{tel que} \quad \ell_i \neq 0$$

D'où la justification d'une partie de l'hypothèse H10.

Th4 : Toute fonction quadratique des y_i , $i = 1...N$,

$$Q = \ell_o + \sum_{i=1...N} (\ell_i y_i + q_{ii} y_i^2) + \sum_{i \neq j=1...N} q_{ij} y_i y_j$$

est estimable, pour un plan de sondage P donné, si et seulement si :

- $\alpha_i > 0$ pour tout i tel que $\ell_i^2 + q_{ii}^2 > 0$
- $\alpha_{ij} > 0$ pour tout $i \neq j$ tel que $q_{ij} + q_{ji} \neq 0$

Donc la variance et le risque ne sont pas estimables pour les plans de sondage en grappe, i.e : tels que :

$$\forall s_1 \neq s_2 \in S \text{ , si } P(s_1) > 0 \text{ et } P(s_2) > 0$$
$$\text{alors } s_1 \cap s_2 = \emptyset$$

dont un des plus usuels est le plan systématique (II-4)

Un résultat intéressant, en faveur de notre restriction aux estimateurs T de y_m de la forme :

$$T = \sum_{k=1...L} M_k T_k$$

est :

C24 : Si pour tout k, T_k ne dépend de S qu'à travers S_k , alors T est P-sans biais si et seulement si, pour $k = 1...L$, le biais de T_k par rapport à P_k est constant, noté t_{ko} (peut être nul) et

$$\sum_{k=1...L} M_k t_{ko} = 0$$

Dem : De 1-b nous avons :

$$T(D) = \sum_{k=1...L} M_k T_k(D_k) \quad \forall D \in \Omega$$

E_p étant un opérateur linéaire, et T_k ne dépendant de S qu'à travers S_k, nous avons :

$$E_p T(D) = \sum_{k=1...L} M_k E_p T_k(D_k)$$

$$= \sum_{k=1...L} M_k E_{p_k} T_k(D_k) \quad \forall D \in \Omega$$

$$d'où \quad B(y) = \sum_{k=1...L} M_k B(yk) \quad \forall y \in R^N$$

avec : $B(y)$ le biais de T, et $B(y_k)$ le biais de T_k , i . e :

$$\forall\ y\ \in R^N \qquad B(y) = E_p\ T(D) - y_m \quad \forall\ y\ \in R^N$$

$$\forall\ y_k \in R^{Nk} \qquad B(y_k) = E_{p_k}\ T_k(D_k) - \bar{y}_k \quad \forall\ y_k \in R^{Nk}$$

La condition suffisante est alors évidente. Montrons qu'elle est nécessaire :

Supposons que T soit P-sans biais, alors :

$$(1)\ O = \sum_k\ M_k\ B(y_k) \qquad \forall\ y \in R^N$$

Supposons qu'il existe au moins un ℓ tel que $B(y_\ell)$ ne soit pas constante. (1) étant vraie pour tout y de R^N, elle est en particulier vraie pour :

- y tel que toutes ses composantes soient nulles, alors :

$$O = \sum_k\ M_k\ B(O_k)$$

où O_k est le point de R^{NK} de composantes toutes nulles.

- y tel que seules les composantes de y_ℓ soient non nulles, alors :

$$O = \sum_{k:k\neq\ell}\ M_k\ B(O_k) + M_\ell\ B(y_\ell) \qquad \forall\ y_\ell \in R^{N\ell}$$

d'où, en combinant les 2 cas :

$$B(y_\ell) = B(O_\ell) \qquad \forall\ y_\ell \in R^{N\ell}$$

en contradiction avec notre hypothèse. D'où C24.

b) La justification des hypothèses H6 et H7 (nous ne considérons que les plans de sondage non ordonnés et sans remise) est dûe à Basu (1969) :

<u>Th5</u> : Soit P un plan de sondage ordonné et avec remise. Et soit s l'échantillon prélevé. Notons s' l'échantillon non ordonné et sans remise déduit de s, i.e. s' est l'ensemble des labels distincts figurants dans s. Alors $d = \{(i, y_i)\ ,\ i \in s'\}$ est exhaustive minimale pour le vecteur y.

D'après le théorème de RAO-BLACKWELL, si P est un plan de sondage ne vérifiant pas H6 ou H7, et si T est un estimateur quelconque de $\bar{y}_k$ *, alors* $E_p(T/d)$ *(d définie en Th3) est un estimateur meilleur que T.*

Il suffit donc de restreindre notre recherche aux estimateurs fonctions de s et de y seulement à travers d. D'où les hypothèses H6 et H7.

c) Dans la suite, nous nous intéresserons particulièrement aux estimateurs P-sans biais.

Sans essayer de justifier un tel choix, nous notons que les estimateurs sans biais présentent, à part leurs qualités intuitives, la propriété d'additivité, définie par GODAMBE et JOSHI (1965) comme suit :

Soient e_r *des estimateurs P-sans biais d'un certain total* z_r *sur une population r, r=1...N... alors* $\sum_{r=1...N} e_r$ *est un estimateur P-sans biais de* $\sum_{r=1...N} z_r$ *, pour tout N. Si nous supposons tous les* e_r *indépendants, de variance bornée, alors par la loi des grands nombres :*

$$\sum_{r=1...N} e_r \Big/ \sum_{r=1...N} z_r \xrightarrow[N \to \infty]{P} 1$$

La convergence étant plus rapide si les variances sont plus petites.

Cependant, si ceci montre l'intérêt qu'on a de considérer, en première approche, des estimateurs sans biais pour estimer la moyenne ou le total sur une population, et de se baser sur la minimisation de la variance comme critère de choix, ça n'exclut nullement la possibilité de trouver des estimateurs biaisés meilleurs que d'autres sans biais.

d) Pour toute donnée $d = \{(i, y_i),\ i \in s\}$ *de* Ω*, et tout y' de* R^N*, notons L(d/y') la fonction de vraisemblance de d sachant y' ([2] , p.45).*

$$L(d/y') = P(s) \prod_{i=1}^{N} 1_{y_i = y'_i}$$

$$\text{avec } 1_{a=b} = \begin{cases} 1 & si \quad a=b \\ 0 & si \quad a \neq b \end{cases}$$

Nous constatons que :

- *L est indépendante du schéma d'échantillonnage engendrant P. Ainsi, si plusieurs schémas engendrent le même plan P, le choix de l'un deux n'influence nullement l'information recueillie.*

- *L donne une information triviale : toutes les composantes non observées de y peuvent prendre toutes les valeurs possibles avec la même probabilité. Ce qui montre les limites du modèle à population fixe et explique des résultats donnés plus tard, tels que la non existence d'estimateur optimal.*

Cependant, le fait qu'il n'y ait aucune hypothèse sur la distribution ξ *ne fait que rendre les résultats plus précieux.*

B-III-3) Modèle de superpopulation

a) Soit M un modèle donné, et soit ξ une distribution de M.

Pour tout échantillon s de S soient :

- ξ_s la distribution marginale de $\{y_i , i \in s\}$
- $\xi_{\bar{s}}$ la distribution conditionnelle des $\{y_i , i \in \bar{s}\}$ étant donné $\{y_i , i \in s\}$

avec $s = \{i_1 \ldots i_n\}$ $\quad i_1 < i_2 < \ldots < i_n$

$\bar{s} = \{j_1 \ldots j_{N-n}\}$ $\quad j_1 < j_2 < \ldots < j_{N-n}$

(voir II-1-b)

Pour toute fonction F(Y) définie sur Ω, notons ($F(Y) = \bar{Y}$ ou $F(Y) = y_m$) :

$$E_M\, T(D/_{S=s}) = \int_{R^N} t(d)\, d\, \xi_s$$

$$E_M\, T(D) = \int_{R^N} T(D/_{Y=y})\, d\, \xi(y)$$

$$V_M\, T(D) = E_M \left(T(D) - E_M\, T(D)\right)^2$$

$$R(P,T) = E_M\, E_p \left(T(D) - F(Y)\right)^2$$

Un estimateur T de F(Y) est ξ - sans biais si, pour tout échantillon s de S :

$$E_M \left(T(D/_{S=s}) - F(Y)\right) = 0$$

T est P ξ - sans biais si :

$$E_p\, E_M \left(T(D) - F(Y)\right) = 0$$

R15 : Les plans de sondage étant non informatifs (H8), les opérateurs E_p et E_M sont commutatifs. D'où si T est P - sans biais (resp. ξ - sans biais) alors il est P ξ - sans biais.

R16 : Dans ce qui précède, nous avons omis volontairement l'indice âge attaché à M, P, ξ, S, Y ...

R17 : *En général, il n'y a aucune raison pour qu'une stratégie ou un plan, retenu dans le cas de l'estimation de* $\bar{Y}$, *présente les mêmes propriétés (optimalité, admissibilité) quand il s'agit d'estimer* $E_M \bar{Y}$. *Nous cherchons à estimer la valeur moyenne* $\bar{Y}$ *sur un verger donné, et non la moyenne sur plusieurs vergers* $E_M \bar{Y}$

Dans ce modèle, la propriété de sans biais peut être requise vis-à-vis du plan de sondage, de la distribution ξ, *ou encore des deux.*

T étant choisi suivant un critère basé sur (P, ξ*), il n'y a aucune raison a priori de privilégier l'un par rapport à l'autre.*

Le cas où T est P ξ *- sans biais est très peu cité dans la littérature, sauf si nous cherchons à estimer* $E_M \bar{Y}$. *En général, on est amené à effectuer un choix :*

D'une part, si nous considérons plusieurs populations provenant d'une même "superpopulation", il semble raisonnable de maintenir la propriété du sans biais comme dans le cas classique (une seule population).

D'autre part, la richesse du modèle de superpopulation provient du fait que nous utilisons nos connaissances a priori pour déterminer, étant donné les observations $\{y_i, \varepsilon s\}$, *la valeur de* $\sum_{i \notin s} y_i$ *et donc de :*

$$\bar{y} = \frac{1}{N} \left[\sum_{i \varepsilon s} y_i + \sum_{i \notin s} y_i \right]$$

ξ *joue donc un rôle vital. Mais aussi, le choix du* ξ *- sans biais suppose que le modèle est "fiable".*

R18 : *le modèle de superpopulation peut aussi être utilisé dans le cas où il y a des erreurs de mesure. L'hypothèse H5 (toute mesure est effectuée sans erreur) exclut cette approche.*

Nous avons considéré les deux approches, en accordant une importance particulière au ξ *- sans biais. La raison découle en partie d'un point de vue pratique pour notre problème : l'estimation de l'effectif moyen, par arbre, sur un verger de pommiers, des oeufs d'hiver de l'araignée rouge. En*

effet, la taille du pommier est en général effectuée à la fin de l'hiver. Il convient de prélever l'échantillon à la même période, de façon à ne tenir compte que des oeufs qui pourraient éclore au printemps, c'est-à-dire non drainés par la pluie. Pour des raisons de coût, il est très probable que l'échantillon soit prélevé seulement parmi les branches qui seront taillées. Le biais, par rapport au plan de sondage, ainsi introduit est variable avec la personne effectuant les prélévements, la méthode de taille, la structure de l'arbre et du verger considéré.

B-III-4) Fonction coût :

Pour déterminer le coût total C(P) du prélévement d'un échantillon suivant le plan de sondage P, il est nécessaire d'avoir une idée sur les couûts relatifs à chacune des opérations :

- travail à accomplir pour prélever l'échantillon : C_T
- Dommages causés au verger par le prélévement : C_D
- Mesures et calculs effectués à partir de l'échantillon : C_M

Nous pouvons considérer que C_M ne dépend que de la taille de l'échantillon sur la population totale. En effet, la méthode utilisée pour évaluer le nombre d'oeufs d'hiver de l'araignée rouge sur un rameau donné, est la même quelque soit son âge : le rameau est passé sous un puissant jet d'eau, et les oeufs sont récupérés dans un filtre facilitant le dénombrement. Donc $C_M = n \cdot C_{Mo}$

Quant à C_T, le fait que l'échantillon soit complétement déterminé avant de se rendre au verger (le plan est non informatif), nous permet de supposer que C_T s'écrit ainsi :

$$C_T = C_{To} + n'_L \cdot C_{T_1} \qquad \text{avec :}$$

- n'_L : le nombre d'arbres différents sur lesquels on effectue au moins un prélévement.

- C_{T_1} : le coût du travail nécessaire pour effectuer des prélévements sur un même arbre, supposé constant quelque soit le nombre de ces prélévements.

- C_{T_O} : un coût minimal donné.

Les dommages causés au verger par le prélévement d'un échantillon $s = \bigcup_{k=1...L} s_k$ dépend de :

- la taille de chacun des échantillons s_k , $k=1...L$

Si C_k est l'évaluation du dommage causé par le prélévement d'un rameau d'âge k, la contrainte d'inclusion H9 nous permet d'affirmer que : $C_k > m_{k-1}\ C_{k-1}$ pour $k = 2...L$

- La dispersion de chacun des échantillons s_k :

En effet, prélever deux rameaux d'âge k sur le même arbre et sur deux arbres différents n'a pas les mêmes conséquences, surtout quand il s'agit de rameaux assez vieux, et a fortiori de plus de 2 rameaux.

Remarquons que par définition des M_k, nous avons :

$$\forall k \ , \ C_k > m_{k-1} \ C_{k-1} \iff M_k \ C_k > M_{k-1} \ C_{k-1}$$

Similairement à M_k , β_k et γ_k , nous définissons (M_k) telle que :

$$\forall k \quad \eta_{k-1} \ M_k \ C_k = \eta_k \ M_{k-1} \ C_{k-1}$$

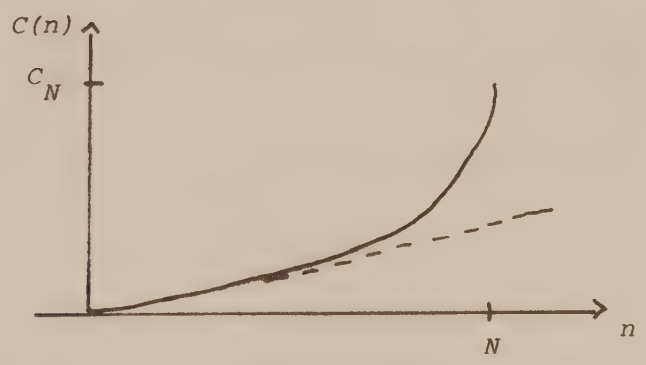
$(\eta_k)_k$ est positive croissante, et : $\eta_L = 1$

nous avons : $\dfrac{\eta_{k-1}}{m_{k-1}} = \dfrac{\eta_k \quad C_{k-1}}{C_k} < 1$

de proche en proche, nous aboutissons à :

$$\eta_1 \ M_k \ C_k = \eta_k \ M_1 \ C_1$$

Dans le cas classique, (une seule population sans stratification), la fonction coût peut être représentée par la courbe



où C_N est le coût total quand on prélève tous les individus (recensement).

Si nous supposons que la taille de l'échantillon est négligeable devant celle de la population, alors la fonction coût peut être approximée par c.n, avec c le coût du prélévement d'un individu.

Nous supposerons désormais que les tailles des différents échantillons s_k , k = 1...L, sont négligeables devant celles des populations.

Dans le cas d'un plan P telle que la dispersion soit maximale, (nous ne prélevons pas deux rameaux de même âge k portés par le même arbre, sauf si un rameau porteur a été prélevé, pour k = 1...L-1) alors, si les tailles des échantillons sont négligeables :

$$C_D(P) = C_L \cdot n_L + \sum_{k=1...L-1} C_k(n_k - m_k\, n_{k+1})$$

or :

$$n_k - m_k\, n_{k+1} = M_k(N_L - n_L)\ (\beta_{k+1} - \beta_k) \qquad (C13)$$

d'où :

$$C_D(P) = \frac{M_1\ C_1}{\eta_1}\ \left[n_L + (N_L - n_L) \sum_{k=1...L-1} \eta_k(\beta_{k+1} - \beta_k)\right]$$

remarquons que dans ce cas, $\delta_k = \beta_k/\beta_{k+1}$ *, d'où :*

$$C_D(P) = \frac{M_1\ C_1}{\eta_1}\ \left[n_L + (N_L - n_L) \sum_{k=1...L-1} \beta_{k+1}\ \eta_k(1 - \delta_k)\right]$$

Nous prendrons, pour tout plan de sondage P, telles que les tailles des échantillons soient négligeables comparées à celles des populations :

$$C_D(P) = \frac{M_1\ C_1}{\eta_1}\ \left[n_L + (N_L - n_L) \sum_{k=1...L-1} \eta_k\ \beta_{k+1}(1 - \delta_k)\right]$$

avec $C_L = M_1\ C_1/\eta_1$ *: dommages causés en prélevant un arbre.*

La détermination des $(\eta_k)_{k=1...L}$ *est faite par récurrence :*

$$\eta_L = 1 \quad \text{et pour} \quad k = L-1...1$$

- *si nous considérons que les dommages sont équivalents dans les cas où la branche d'âge k+1 est prélevée et ou si seulement tout ce qu'elle porte (les branches d'âge k) est prélevé alors* $\eta_k = \eta_{k+1}$

- *sinon, donc si le rameau d'âge k, tout seul, présente un intérêt particulier (par rapport aux fruits ou à la croissance...) alors :*

$$\eta_k < \eta_{k+1} \quad \text{et} \quad \frac{\eta_k}{\eta_{k+1}}$$

traduit l'importance de l'âge considéré.

La fonction coût total est :

$$C(P) = C_M(P) + C_T(P) + C_D(P)$$

C_T et C_D variant inversement par rapport à la dispersion du plan de sondage, alors que C_M ne dépend que du nombre total de tous les rameaux prélevés.

Finalement, nous remarquons que par la nature du problème (contrainte d'inclusion), la fonction coût ne peut être définie que pour le plan de sondage P sur la population totale.

BIBLIOGRAPHIE :

[1] *BASU, D. (1969)*
Role of the sufficiency and Likelihood principles in sample survey theory
Sankhya, A31, p. 441-453

[2] *CASSEL, C.M. ; SARNDAL, C.E. ; WRETMAN, J.H. (1977)*
Foundations of inference in survey sampling
John Wiley & Sons

[3] *COCHRAN (1962)*
Sampling Techniques - 5ème édition
John Wiley & Sons

[4] *GODAMBE , V.P. ; JOSHI, V.M. (1965)*
Admissibility and Bayes estimation in sampling finite population, I.
Ann. Math. Statist., 36, p. 1707-1722

[5] *HANURAV, T.V. (1966)*
Some aspects of unified sampling theory
Sankhya, A28, p. 175-204

[6] *ROYALL, R. (1968)*
An old approach to finite population sampling theory
J. Amer. Stat. Ass., 63, p. 1269-1279

[7] *ROYALL, R. (1976)*
Likelihood functions in finite population sampling theory
Biometrika, 63, p. 605-614

[8] *YATES, F. (1981)*
Sampling methods for censuses and Surveys - 4ème édition
C. GRIFFIN SCLTD

B - IV - MODELE A POPULATION FIXE

Si aucune hypothèse n'est faite sur les valeurs prises par la variable mesurée, alors nous n'avons pas de stratégie optimale. Dans 1-b), nous choisissons la stratégie (P, T_{gd}) qui vérifie des propriétés intéressantes résumées dans Th7.

Dans 2) nous appliquons cette stratégie à notre cas, et discutons le choix d'un estimateur du risque.

La complexité du problème ne nous a pas permis d'aller au bout, pour la détermination des choix optimaux des paramètres $\{\gamma_k , \delta_k\}_{k=1...L-1}$

B-IV-1) Choix d'une stratégie

a) Non existence d'estimateurs optimaux et de stratégies optimales

Considérons un plan de sondage P, vérifiant les hypothèses H6, H7, H8 et telles que ses probabilités d'inclusion d'ordre 1 vérifient :

- $\forall i \in U \quad \alpha_i > o$ (Th3 - H10)
- $\exists i \in U$ tel que $\alpha_i < 1$ (P n'est pas un recensement)

Etant donné une classe C d'estimateurs de $\bar{y}$, existe-t-il un estimateur de C dont le risque soit inférieur, pour tout y de R^N, au risque de tout autre élément de C ?

Dans le cas où C est la classe de tous les estimateurs T de $\bar{y}$, sans conditions sur T, ou encore dans le cas où C est la classe de tous les estimateurs linéaires de $\bar{y}$, la réponse est non.

En effet, supposons qu'il existe un estimateur (linéaire) To de $\bar{y}$ tel que :

$$E_p(To - \bar{y})^2 \leq E_p(T - \bar{y})^2$$

pour tout estimateur T (linéaire) de $\bar{y}$ et pour tout y de R^N. Donc, en particulier pour T égal à une constante co et y tel que $\bar{y} = co$, nous devons avoir $E_p(To - \bar{y})^2 = o$. Ceci étant vrai pour tout choix de co, nous avons :

$$E_p(To - \bar{y})^2 = O \quad \forall y \in R^N$$

ce qui n'est possible que si :

$$\forall s \in S , \ \forall y \in R^N \quad to(d) = \bar{y}$$

Or to ne dépend de $y,\ldots y_N$ qu'à travers les y_i tels que $i \in s$. Donc $s = U$, ou encore que le plan de sondage est un recensement ce qui est en contradiction avec l'hypothèse :

$$\exists i \in U \quad \text{tel que} \quad \alpha_i < 1$$

En particulier, nous en déduisons que toute constante est admissible parmi tous les estimateurs, linéaire ou non, de $\bar{y}$.

Si maintenant la propriété de P-sans biais est requise, alors GODAMBE et JOSHI (1965) ont montré qu'il n'existe pas d'estimateur optimal dans la classe des estimateurs P-sans biais (resp. linéaires P-sans biais) de $\bar{y}$.

Ce résultat reste valable si nous ne considérons que les estimateurs P-sans biais linéaires homogènes de $\bar{y}$ sauf dans le cas où P est un plan en grappe (voir III-2). En effet, pour de tels plans, l'estimateur de HORVITZ-THOMPSON :

$$T_{HT}(D) = \sum_{i \varepsilon s} y_i / \alpha_i$$

est l'unique estimateur linéaire homogène et P-sans biais de $\bar{y}$ ([1], p.70).

Il découle de ce qui précède la non existence de stratégie optimale parmi toutes les stratégies (P , T) telles que :

- P n'est pas un recensement et $\forall i \varepsilon U \quad \alpha_i > 0$
- T est un estimateur (resp. linéaire, P sans-biais, ou linéaire P-sans biais) de $\bar{y}$.

En effet, sinon il existerait (Po, To) meilleure que toute autre stratégie d'une classe C , et en particulier (Po,To) est meilleure que toutes les stratégies (Po, T) où T appartient à la classe d'estimateur correspondante C. Donc sous Po, To est optimal dans C.

b) Stratégie (P, Tgd)

Bien des estimateurs sont admissibles dans la classe des estimateurs P-sans biais de $\bar{y}$. Nous allons nous intéresser particulièrement à l'estimateur de la différence généralisée, qui présente d'autres propriétés intéressantes.

Soit e un point arbitraire de R^N - l'estimateur de la différence généralisée associée est défini par :

$$\forall d \varepsilon \Omega \quad t_{gd}(d) = \bar{e} + \sum_{i \varepsilon s} \frac{y_i - e_i}{N \alpha_i}$$

l'admissibilité de T_{gd} découle facilement d'un théorème de GODAMBE et JOSHI (1965), que nous généralisons comme suit :

Th6 Toute fonction $g(s , y)$ définie sur $S \times R^N$ qui ne dépend de (s,y) qu'à travers les (i , y_i) tels que $i \in s$ et $y_i \neq e_i$ est un estimateur admissible parmi tous les estimateurs P-sans biais de :

$$G(y) = E_p\, g(s , y) = \sum_{s \in y} g(s , y).P(S=s)$$

GODAMBE et JOSHI ont montré ce théorème dans le cas où :

$$\forall\, i \in U \quad e_i = 0$$

D'autre part, il est facile de constater, comme le font remarquer les auteurs dans le cas $e = 0$, que le Th6 reste valable dans le cas où y est restreint à toute partie de R^N contenant e.

Donc Tgd est en fait admissible parmi tous les estimateurs P-sans biais de $\bar{y}$ même si y ne varie que dans $\prod_{i=1}^{N} [a_i + e_i , b_i + e_i]$ pour tous $(a_i , b_i)_{i=1...N}$ tels que $a_i \geq 0$ et $b_i > 0$.

En fait, une propriété moins forte encore distingue T parmi tous les estimateurs P-sans biais de $\bar{y}$, comme nous allons le voir pour le cas où $e = 0$, la généralisation est aisée en reprenant les démonstrations (voir [1] p.67, rem.4)

HANURAV (1965) définit le critère d'hyper admissibilité d'un estimateur comme suit :

- une hyper surface principale $A(i_1,\ldots,i_r)$ est l'ensemble des points y de R^N tels que $y_j = 0$ $(y_j = e_j)$ pour $j \neq i_1 \ldots i_r$, pour tout sous ensemble $\{i_1 \ldots i_r\}$ non vide de U.

- un estimateur T est dit hyper admissible, dans une classe C d'estimateurs P-sans biais d'une fonction $f(y)$ s'il est admissible dans C quand y est restreint à l'hyper surface principale $A(i_1 \ldots i_r)$ de R^N, pour tout choix de $\{i_1,\ldots,i_r\}$ non vide de U.

LANKE et RAMAKRISHNAN(1974) ont montré que si le plan de sondage n'est pas en grappe, ou s'il est en grappe avec au moins trois grappes, alors

T_{HT} (T_{gd}) est l'unique hyper admissible estimateur dans la classe de tous les estimateurs P-sans biais de $\bar{y}$.

Dans la pratique, si le choix du point arbitraire e de R^N est bien fait, nous savons que y est restreint à une partie de R^N de la forme $\prod_{i=1}^{N} [- a_i + e_i , b_i + e_i]$ avec $a_i \geq 0$ et $b_i > 0$

Par exemple, pour nos données, nous savons que y est positive. Si nous ne considérons que les rameaux de 1 an , nous pouvons affirmer que :

$$\forall i \in U_1 \quad 0 \leq y_{1i} < 10.000$$

D'autre part, pour différents choix des e_i et α_i , $i \in U$, nous retrouvons des estimateurs très utilisés tels que $\bar{y}_s$: la moyenne sur l'échantillon, et T_{HT} : l'estimateur de HORVITZ-THOMPSON.

Toutes ces raisons nous ont poussé à ne considérer dans la suite que l'estimateur T_{gd}. En fait, T_{gd} est défini étant donné un point arbitraire e de R^N , et un plan de sondage P:

- P de probabilités d'inclusion $\{\alpha_i , \alpha_{ij} ; i \neq j \in U\}$ telles que :

$$\forall i \in U \quad \alpha_i > 0$$

- $\forall s \in S$

$$T_{gd}(D/_{S=s}) = t_{gd}(d) = \bar{e} + \sum_{i \in s} \frac{y_i - e_i}{N \alpha_i}$$

Th7 : ([1], p.62)
la stratégie (P, T_{gd}) est admissible parmi toutes les stratégies (P',T') telle que :
- P' est tel que $E_{p'} \nu(S) = E_p \nu(S)$
- T' est P'-sans biais

B-IV-2) Propriétés de la stratégie (P, T_{gd})

a) Etant donnés un point e de R^N, et un plan de sondage des probabilités d'inclusion $\alpha_i > 0$, $i \in U$

Th8 : ([1], p.74)
Si P est de taille fixe n , alors T_{gd} est optimal parmi tous les estimateurs P-sans biais de $\bar{y}$ ne dépendant de (s,y) qu'à travers

$$\{\frac{y_i - e_i}{\alpha_i} : i \in s\}$$

Comment choisir alors e et α ? ce choix devrait nous permettre d'avoir un plus faible risque ou variance (puisque T_{gd} est P-sans biais) :

$$V_p\, T_{gd}(D) = \frac{1}{N^2} \sum_{i \in U} (y_i - e_i)^2 \frac{1 - \alpha_i}{\alpha_i} + \frac{1}{N^2} \sum_{i \neq j \in U} \frac{y_i - e_i}{\alpha_i} \frac{y_j - e_j}{\alpha_j} (\alpha_{ij} - \alpha_i \alpha_j)$$

Si le plan de sondage est à taille fixe, alors :

$$\sum_{\substack{j \in U \\ j \neq i}} \alpha_{ij} - \alpha_i \alpha_j = \alpha_i(\alpha_i - 1) \qquad (Th1)$$

En remplaçant dans $V_p\, T_{gd}$, nous avons :

$$V_p\, T_{gd}(D) = \frac{1}{N^2} \sum_{i < j \in U} (\alpha_i \alpha_j - \alpha_{ij}) \left(\frac{y_i - e_i}{\alpha_i} - \frac{y_j - e_j}{\alpha_j}\right)^2$$

De cette deuxième expression, on voit facilement que si $\frac{y_i - e_i}{\alpha_i}$ est constante pour tout i, alors $V_p\, T_{gd}$ est nulle.

Supposons qu'il existe une variable auxiliaire x très corrélée avec y : $\forall i \in U \quad y_i \overset{\sim}{\sim} a + b\, x_i$ avec a et b connus. Soit n la taille fixe de P. Sans perte de généralités, nous pouvons supposer que x est positive et vérifie :

$$\forall i \in U \quad n.x_i < N \bar{x} \qquad (voir\ II\text{-}3\text{-}a)$$

Nous avons alors principalement deux possibilités :

a) - $\forall i \in U \quad ei = a \quad$ et $\quad \alpha_i = \dfrac{nx_i}{N\bar{X}}$

où a est une constante donnée. Alors

$$T_{gd}(D) = a\left[1 - \frac{\bar{X}}{n} \sum_{i \in S} \frac{1}{x_i}\right] + \frac{\bar{X}}{n} \sum_{i \in S} \frac{y_i}{x_i}$$

En particulier, si a=o, nous retrouvons l'estimateur très connu :

$$\frac{\bar{X}}{n} \sum_{i \in S} \frac{y_i}{x_i}$$

Si la variable X est qualitative, alors prendre α_i proportionnelle à x_i signifie que les α_i sont constantes pour tous les i d'une même strate (déterminée par un niveau de X) et nous retrouvons l'estimateur de HORVITZ-THOMPSON.

b) - $\forall i \in U \quad ei = b.x_i \quad$ et $\quad \alpha_i = \dfrac{n}{N}$

alors :

$$T_{gd}(D) = \bar{y}_S + b(\bar{X} - \bar{X}_S)$$

qui est un autre estimateur très connu.

- Dans le cas où e et α sont proportionnels à X, nous retrouvons le 1er cas.

- Signalons une troisième possibilité :

$\forall i \in U \quad e_i = b \cdot X_{1i} \quad \alpha_i = \dfrac{n X_{2i}}{N \bar{X}_2}$

$$\text{Alors } T_{gd} = b(\bar{X}_1 - \sum_{i \in s} \frac{X_{1i}}{N\alpha_i}) + \sum_{i \in s} \frac{y_i}{N\alpha_i}$$

D'après les résultats de l'analyse de covariance, avec comme covariable la longueur, cette possibilité peut présenter des avantages. Si X_1 est la longueur, et X_2 la variable qualitative correspondant à la stratification, $\dfrac{y - ei}{N\alpha_i}$ a une plus faible variabilité que $\dfrac{y_i}{N\alpha_i}$ ou y_i. Il faut cependant connaitre la longueur moyenne de tous les rameaux du verger ainsi que le coefficient de régression b (éventuellement les coefficients de régression pour les différents âges).

Dans la suite, nous supposons que e=o. La démarche est similaire si e ≠ o.

Pour l'estimation de $\bar{y}$: la moyenne sur tous les rameaux du verger, nous considérons donc la stratégie (P, T_{HT}), avec :

- P comme décrit en II, de probabilités d'inclusion

$$\alpha_i,\ \alpha_{ij} : i \neq j \in U$$

- T_{HT} défini par :

$$T_{HT}(D) = \frac{1}{N} \sum_{i \in S} \frac{y_i}{\alpha_i} \quad \forall y \in R^N$$

D'après R1, la stratégie (P,T) pour l'estimation de y_m, où T est défini par :

$$T(D) = \frac{1}{N_L} \sum_{i \in S} \frac{y_i}{\alpha_i} = \sum_{k=1 \ldots L} M_k \cdot T_{HT}(D_k)$$

jouit des mêmes propriétés que (P, T_{HT}) pour l'estimation de $\bar{y}$.

Nous en déduisons :

$$V_p\left(T(D)\right) = \frac{1}{N_L{}^2} \left[\sum_{i \in U} y_i^2 \frac{1 - \alpha_i}{\alpha_i} + \sum_{i \neq j \in U} y_i y_j \frac{\alpha_{ij} - \alpha_i \alpha_j}{\alpha_i \alpha_j} \right]$$

$$= \frac{1}{N_L{}^2} \left[\sum_{i \in U} (\alpha_i \alpha_j - \alpha_{ij}) \left(\frac{y_i}{\alpha_i} - \frac{y_j}{\alpha_j}\right)^2 \right]$$

Bien sûr V_p T(D) ne peut être déterminée exactement, il n'est estimable que si :

$$\forall i \neq j \in U \quad \alpha_i > 0 \ \text{ et } \ \alpha_{ij} > 0 \qquad (Th4)$$

Ce qui exclut tous les plans en grappe.

Deux estimateurs P-sans biais de V_p T ont été avancés. Le premier par HORVITZ et THOMPSON (1952) :

$$\forall s \in S \quad v(d) = \frac{1}{N_L{}^2} \left[\sum_{i \in s} (1 - \alpha_i) \left(\frac{y_i}{\alpha_i}\right)^2 + \sum_{i \neq j \in s} \frac{\alpha_{ij} - \alpha_i \alpha_j}{\alpha_{ij}} \frac{y_i}{\alpha_i} \frac{y_j}{\alpha_j} \right]$$

et le second par YATES et GRUNDY (1953) :

$$\forall s \in S \quad v'(d) = \frac{1}{N_L{}^2} \sum_{i \neq j \in s} \frac{\alpha_i \alpha_j - \alpha_{ij}}{\alpha_{ij}} \left(\frac{y_i}{\alpha_i} - \frac{y_j}{\alpha_j} \right)^2$$

En général, v et v' sont distincts. Un exemple simple, dans le cas d'une population classique, est si $s = \{ i \}$ *: seul l'individu i a été prélevé :*

$$v = \frac{1}{N^2} (1 - \alpha_i) \left(\frac{y_i}{\alpha_i} \right)^2 \quad \text{alors que} \quad v' = 0$$

D'autre part, le Th6 montre que v est un estimateur admissible parmi tous les estimateurs P-sans biais de $V_p(T(D))$. *En effet, v ne dépend de* (s , y) *qu'à travers les couples* (i , y_i) *tels que* $i \in s$ *et* $y_i \neq 0$. *Nous en déduisons la non existence d'estimateur optimal de* $V_p(T(d))$.

Finalement, nous signalons un argument en faveur de H_{11} :

<u>Th9</u> : [5]
Pour tout plan de sondage P à taille fixe n, de probabilité d'inclusion α_i , $i \in U$, *il est impossible de construire un plan de sondage P' de taille aléatoire et ayant les mêmes probabilités d'inclusion d'ordre 1, tel que :*

$$V_p , (v(d)) \leqslant V_p (v(d)) \quad \forall \quad y \in R^N$$

b) L'expression de $V_p\ T(D)$ *en fonction des probabilités d'inclusion de la population* U_L *et des paramètres* $\{\beta_k , \gamma_k \ \delta_k\}_{k=1...L-1}$ *est très compliquée. Nous avons longtemps cherché à l'exprimer simplement, ainsi qu'à déterminer le choix optimal des* $\{\gamma_k , \delta_k\}_{k=1...L-1}$, *mais sans succès.*

Le choix des paramètres $\{\beta_k , \gamma_k , \delta_k\}$ *est naturellement fonction de la répartition des valeurs* y_i , $i \in U$. *Il aurait été intéressant de confirmer des notions intuitives telles que : si, pour un âge donné k, la variance inter arbre des rameaux de* U_k *est faible comparée à la variance intra arbre, alors prendre* δ_k *maximal et vice-versa. Ceci n'est nullement impossible, et la démarche devrait être comparable au cas d'une population classique stratifiée pour la détermination des tailles optimales (allocation de NEYMAN) Nous présentons néanmoins quelques résultats.*

Notons : $\forall k \leqslant \ell$, $\forall i \in U_\ell$,

$$B_k(\ell , i) = \sum_{j \in W_i \cap U_k} y_{kj}$$

$$V_p\ T(D) = \frac{1}{N_L{}^2} \left[\sum_{i \in U} \frac{1}{\alpha_i} y_i^2 + \sum_{i \neq j \in U} \frac{\alpha_{ij}}{\alpha_i \alpha_j} y_i\ y_j \right] - y_m^2$$

Ecrivons : $\forall\, i \neq j \in U \qquad \alpha_{ij} = (1-\alpha_i-\alpha_j+\alpha_{ij}) - (1-\alpha_i-\alpha_j)$

$$V_p\, T(D) = \frac{1}{N_L{}^2}\left[\sum_{i\in U} y_i^2\,\frac{1-\alpha_i}{\alpha_i^2} - \Big(\sum_{i\in U}\frac{y_i}{\alpha_i}\Big)^2 + \sum_{i\neq j\in U}\frac{1-\alpha_i-\alpha_j+\alpha_{ij}}{\alpha_i\alpha_j}\,y_i\,y_j\right]$$

$$+\ \frac{1}{N_L}\,y_m\sum_{i\in U} y_i\,\frac{2-\alpha_i}{\alpha_i}$$

Le seul terme où interviennent γ_k *et* δ_k, $k=1\ldots L$, *est :*

$$\sum_{i\neq j\in U}\frac{1-\alpha_i-\alpha_j+\alpha_{ij}}{\alpha_i\,\alpha_j}\,y_i\,y_j \qquad\qquad (C12 - C21)$$

Nous le décomposons, suivant l'âge et l'arbre porteur, en la somme des deux termes (C5) :

$$A = \sum_{i\neq i'\in U_L}\ \sum_{h,k=1\ldots L}\ \sum_{\substack{j\in W_i\cap U_k\\ j'\in W_{i'}\cap U_h}}\frac{1-\alpha_j-\alpha_{j'}+\alpha_{jj'}}{\alpha_j\,\alpha_{j'}}\,y_{kj}\,y_{hj'}$$

$$B = \sum_{i\in U_L}\ \sum_{h,k=1\ldots L}\ \sum_{\substack{j\in W_i\cap U_k\\ j'\in W_i\cap U_h\\ j\neq j'}}\frac{1-\alpha_j-\alpha_{j'}+\alpha_{jj'}}{\alpha_j\,\alpha_{j'}}\,y_{kj}\,y_{hj'}$$

En tenant compte de la hiérarchie des âges, et en appliquant C21-i, nous avons :

$$A = \sum_{i\neq i'\in U_L}(1-\alpha_i-\alpha_{i'}+\alpha_{ii'})\left[\sum_{k=1}^{L}\gamma_k\,\frac{B_k(L,i)}{1-(1-\alpha_i)\beta_k}\,\frac{B_k(L,i')}{1-(1-\alpha_{i'})\beta_k}\right.$$

$$\left.+\ 2\sum_{h<k}\beta_h\,\frac{B_h(L,i)}{1-(1-\alpha_i)\beta_h}\,\frac{\gamma_k}{\beta_k}\,\frac{B_k(L,i')}{1-(1-\alpha_{i'})\beta_k}\right]$$

Pour B, nous considérons les couples de rameaux portés par le même arbre. Il faut alors tenir compte de l'âge du plus jeune rameau porteur des 2 rameaux considérés (C5 - ii).

En décomposant suivant toutes les possibilités, et en appliquant H9 pour des rameaux portés l'un par l'autre, et C21-ii pour les autres cas, nous avons :

$$B = \sum_{i\in U_L} (1-\alpha_i) \sum_{h<k} \sum_{j\in W_i\cap U_k} 2\,\beta_h \frac{y_{kj}}{1-(1-\alpha_i)\beta_k} \frac{B_h(k,j)}{1-(1-\alpha_i)\beta_h}$$

$$+ \sum_{i\in U_L} (1-\alpha_i) \sum_{h<k<\ell} \sum_{r\in W_i\cap U_\ell} \sum_{\substack{j\neq j'\\ \in W_r\cap U_{\ell-1}}} 2\,\frac{\beta_h\gamma_k\delta_{\ell-1}\beta_\ell}{\beta_k\,\gamma_{\ell-1}} \frac{B_k(\ell-1,j)}{1-(1-\alpha_i)\beta_k} \frac{B_h(\ell-1,j')}{1-(1-\alpha_i)\beta}$$

$$+ \sum_{i\in U_L} (1-\alpha_i) \sum_{h<k} \sum_{r\in W_i\cap U_k} \sum_{\substack{j\neq j'\\ \in W_r\cap U_{k-1}}} \frac{\gamma_h\delta_{k-1}\beta_k}{\gamma_{k-1}} \frac{B_h(k-1,j)\,B_h(k-1,j')}{[1-(1-\alpha_i)\beta_h]^2}$$

Remarquons alors que :

$$\forall i \in U_L \quad , \quad \forall h < k < L$$

$$\sum_{\ell>k} \sum_{r\in W_i\cap U_\ell} \sum_{j\neq j'\in W_r\cap U_{\ell-1}} \frac{\delta_{\ell-1}\,\beta_\ell}{\gamma_{\ell-1}} B_k(\ell-1,j)\,B_h(\ell-1,j')$$

$$= \sum_{\ell=k+1}^{L} \frac{\delta_{\ell-1}\,\beta_\ell}{\gamma_{\ell-1}} \left[B_k(\ell,r)\,B_h(\ell,r) - \sum_{j\in W_r\cap U_{\ell-1}} B_k(\ell-1,j)\,B_h(\ell-1,j) \right]$$

$$= \sum_{\ell=k}^{L-1} \left(\frac{\delta_{\ell-1}\,\beta_\ell}{\gamma_{\ell-1}} - \frac{\delta_\ell\,\beta_{\ell+1}}{\gamma_\ell} \right) \sum_{r\in W_i\cap U_\ell} B_k(\ell,r)\,B_h(\ell,r)$$

$$+ \frac{\delta_{L-1}}{\gamma_{L-1}} B_k(L,i)\,B_h(L,i) - \frac{\delta_{k-1}\,\beta_k}{\gamma_{k-1}} \sum_{r\in W_i\cap U_k} y_{kr}\,B_h(k,r)$$

Dans le cas particulier (II-4-ii) où :

$$\forall\ k = 1\ldots L-1 \qquad \frac{\delta_k\,\beta_{k+1}}{\gamma_k} = \frac{\delta_{L-1}}{\gamma_{L-1}}$$

cette dernière expression devient :

$$\frac{\delta_{L-1}}{\gamma_{L-1}} \left[B_k(L,i)\,B_h(L,i) - \sum_{r\in W_i\cap U_k} y_{kr}\,B_h(k,r) \right]$$

D'où l'expression de B dans ce cas :

$$B = \frac{\delta_{L-1}}{\gamma_{L-1}} \sum_{i \in U_L} (1-\alpha_i) \Big\{ 2 \sum_{h<k} \frac{\beta_h \gamma_k}{\beta_k} \frac{B_k(L,i)}{1-(1-\alpha_i)\beta_k} \frac{B_h(L,i)}{1-(1-\alpha_i)\beta_h}$$

$$+ 2 \sum_{h<k} \beta_h \left(\frac{\gamma_{L-1}}{\delta_{L-1}} - \frac{\gamma_k}{\beta_k}\right) \sum_{j \in w_i \cap U_k} \frac{y_{kj}}{1-(1-\alpha_i)\beta_k} \frac{B_h(k,j)}{1-(1-\alpha_1)\beta_h}$$

$$+ \sum_h \gamma_h \frac{B_h^2(L,i) - \sum_{j \in w_i \cap U_h} y_{hr}^2}{\left[1-(1-\alpha_i)\beta_h\right]^2} \Big\}$$

B I B L I O G R A P H I E

[1] C.M. CASSEL - C.E. SARNDAL - J.H. WRETMAN (1977)

Foundations of inference in survey sampling
Wiley - Interscience

[2] W.G. COCHRAN (1962)

Sampling Techniques
John Wiley & Sons

[3] V.P. GODAMBE - V.M. JOSHI (1965)

Admissibility and bayes estimation in sampling finite population. I.
Ann. Math. Stat. 36, p. 1707-1722

[4] T.V. HANURAV (1968)

Hyper-admissibility and optimum estimators for sampling finite populations
Ann. Math. Stat. 39, p. 621-642

[5] V.M. JOSHI (1965)

Admissibilité and bayes estimation in sampling finite population.
II-III Ann. Math. Stat. (36) p. 1723-1742
IV Ann. Math. Stat. (37) p. 1658-1670

[6] V.M. JOSHI (1969)

Admissibility of estimates of the mean of a finite population.
In. N.L. Johnson & H. SMITH Eds
New developments in survey sampling

[7] J. LANKE, M.K. RAMAKRISHNAN (1974)

Hyperadmissibility in survey sampling
Ann. Stat. 2 p. 205-215

[8] F. YATES (1981)

Sampling methods for censures and survey
4ème édition - C. GRIFFIN SGLTD

B - V - MODELE DE SUPERPOPULATION

Dans ce chapitre, nous traduisons nos connaissances a priori sur la distribution de la variable par la définition d'une classe de distribution à travers un modèle résumant les hypothèses que cette distribution est censée vérifier.

Nous nous sommes intéressés à trois modèles généraux, chacun recouvrant un aspect des résultats de nos données (A-II). Chacune des populations est d'abord considérée séparément, et la stratégie optimale pour l'estimation de $\bar{Y}_k$, $k = 1...L$, est définie. Pour l'estimation de Y_m, la moyenne par arbre, il suffit alors de considérer l'estimateur défini en III.

Dans 1), nous définissons les 3 modèles considérés.

Dans 2), nous nous intéressons aux stratégies P-sans biais de $\bar{Y}_k$, $k=1...L$, sous le modèle M1, et puis mettons le lien avec les résultats de IV.

Dans 3), nous exposons l'esprit général pour déterminer les stratégies ξ-sans biais.

Les parties 3, 4 et 5 répondent alors à ce problème sous les 3 modèles respectivement. Finalement, dans 6 nous définissons l'approche adoptée pour estimer Y_m (la moyenne par arbre sur tout le verger).

D'autres modèles plus compliqués sont plausibles, notamment des modèles résultant d'une combinaison de ceux considérés. Ces cas n'ont pas été traités.

B-V-1) Modèles considérés à partir des traitements des données. (l'indice âge est omis)

Le premier modèle est formulé à partir des résultats de l'analyse de variance (A-II-6). Nous supposons que la répartition de la variable n'est pas uniforme sur tout le verger : il existe des strates (aggrégats, concentrations...) telles que les variances intra strates puissent être très faibles comparées à la variance inter. La formulation du modèle peut cependant correspondre à d'autres problèmes :

Modèle M1 :

Il existe des constantes connues a_i , b_i , pour $i \in U$ tels que :

$$\forall i \in U \quad E_M Y_i = a_i \mu + b_i$$

$$V_M Y_i = a_i^2 \sigma^2$$

$Y_1 \ldots Y_N$ sont indépendantes, μ et σ^2 inconnues, de fonction de distribution absolument continue

Les a_i $i \in U$ sont supposés strictement positifs, de moyenne $\bar{a} = \frac{1}{N} \sum_{i \in U} a_i = 1$.

Pour nos données, pour chacune des populations U_1, U_2, U_3, désignons par j un rameau et i le numéro de sa strate (Rangée x Niveau x face). Le modèle s'applique alors de deux façons :

a) $Log(Y_j) = e_i + \varepsilon_j$

avec ε_j, $j \in U$, indépendantes de même loi $N(0,S^2)$. Alors :

$$E_M Y_j = \exp(e_i) \cdot \exp(\frac{1}{2} S^2)$$

$$V_M Y_j = \exp(2 e_i + S^2) \cdot (\exp S^2 - 1)$$

Nous posons : $a_j = a_i = C \cdot \exp(e_i)$

$$b_j = 0$$

alors $E_M Y_j = C' a_i \exp(\frac{1}{2} S^2) = a_i \mu$

$$V_M Y_j = C'' a_i^2 \exp S^2 (\exp S^2 - 1) = a_i^2 \sigma^2$$

où C, C', C" sont des constantes.

b) $Y_j = e_i + \varepsilon_j$

avec ε_j, $j \in U$, indépendantes centrées, de variance σ_i^2

Ce cas est repris plus en détail dans 2-b. Ici, nous posons :

$$b_j = e_i$$

$$a_j = C \cdot \sigma_i$$

(Ces paramètres peuvent être définis au niveau des individus en tenant compte, par exemple, en même temps de la stratification et des longueurs).

Dans ces deux cas, C est une constante telle que $\bar{a} = 1$. Les coefficients a_j , b_j sont constantes dans la même strate. En fait, c'est ainsi que nous définissons les strates, chacune correspondant à un degré d'attaque.

Nous poserons alors : $Z_j = \dfrac{Y_j - b_j}{a_j}$ $\quad \forall j \in U$

Le modèle M1 suppose donc que les Z_j sont indépendantes, de même moyenne et de même variance, et qu'ils ont des fonctions de distribution absolument continues. Dans le modèle M2, nous supposons que les Z_j sont indépendantes, identiquement distribuées, de fonction de distribution connue.

<u>Modèle M2</u> :

Le modèle M1 est vérifié et :

$$Z_j = \frac{Y_j - b_j}{a_j} , \quad j \in U$$

sont indépendantes, identiquement distribuées, de fonction de distribution $F(z/\theta)$ connue (θ inconnue).

Si pour nos données, nous avons considérés la première approche, alors :

$$\text{Log } Z_j = \text{Log } Y_j - \text{Log } a_j = \varepsilon_j$$

(avec les notations du cas a dans le modèle M1)

Donc les Z_j seraient distribués suivant une loi lognormale.

Indépendamment de ces 2 premiers modèles, nous définissons M3 à partir des résultats de la régression : le nombre d'oeufs d'hiver de P-Ulmi déposés sur un rameau j de U_k , *k=2...L, est corrélé fortement et positivement avec les totaux des dépôts sur tous les rameaux de la branche* W_j *d'âge inférieur à k.*

La contrainte d'inclusion H9 implique que le nombre de branches d'âge k sur lesquelles un rameau d'âge h, h < k, a été prélevé est supérieur au nombre de rameaux d'âge k qui ont été prélevé. Nous allons considérer les données $\{D_h$; $h < k\}$ *comme information auxiliaire, alors qu'en fait elle n'est connue que sur une partie (contenant* S_k*) de la population.*

Modèle M3 :

Il existe une variable auxiliaire X (multidimensionnelle), connue sur toute la population U, telle que :

$$E_M Y_j = X_j \cdot b = \sum_{h=1}^{q} x_{hj} \cdot b_h$$

$$V_M Y_j = \sigma^2 \cdot v_j$$

avec $b \in R^q$ *le vecteur paramètre, inconnu.*

$v_j = v(X_j)$ *où v est une fonction connue.*

$Y_1 \ldots Y_N$ *indépendantes. (éventuellement* X_1 *tel que toutes ses composantes soient égales à 1).*

D'autres modèles peuvent être définis, notamment en utilisant les résultats de l'analyse de covariance (remplacer dans le modèle M3 les Y par les Z tels que définis dans M1). Ce modèle ne devrait pas présenter de difficultés, mais n'a pas été considéré.

B-V-2) Modèle I, stratégies P-sans biais

a) Soit P_O un plan de sondage non ordonné, sans remise, non informatif (H6, H7, H8), de taille fixe n (H11) et tel que ses probabilités d'inclusion α_i soient proportionnelles aux a_i :

$$\forall i \in U \qquad \alpha_i = \frac{n}{N} a_i$$

Notons T_O l'estimateur de la différence généralisée, comme défini en IV, quand e=b :

$$T_O(D) = \bar{b} + \sum_{i \in S} \frac{Y_i - b_i}{n\, a_i}$$

Si la distribution jointe ξ de $Y_1 \ldots Y_N$ vérifie le modèle M1, alors T_O est ξ-sans biais. En effet :

$$E_M\, T_O(D) = \bar{b} + \sum_{e \in S} \frac{1}{n a_i} (E_M\, Y_i - b_i)$$

$$= \bar{b} + \sum_{i \in S} \frac{1}{N}\, \mu_i = \mu + \bar{b}$$

$$\text{et} \quad E_M\, \bar{Y} = E_M\left(\frac{1}{N} \sum_{i \in U} Y_i\right) = \frac{1}{N} \sum_{i \in U} E_M(Y_i)$$

$$= \frac{1}{N} \sum_{i \in U} (a_k\, \mu + b_k) = \bar{b} + \mu$$

T_O est P_O-sans biais, mais peut-être biaisé si nous considérons un autre plan de sondage.

Les propriétés de la stratégie (P_O, T_O) sont résumés dans le TH suivant :

TH10 : : ([2], p. 103)

Si ξ vérifie M1, notons $\Delta = \sum_{i=1}^{N} a_i^2/N$. Alors :

$$R(P,T) \geq R\,(P_O, T_O) = \frac{\sigma^2}{n}\left(1 - \frac{n}{N}\Delta\right)$$

pour toute stratagie (P,T) telle que P soit de taille fixe n de probabilités d'inclusion d'ordre 1 strictement positives, et T un estimateur P-sans biais de $\bar{Y}$. Nous avons égalité si et seulement si $(P,T) = (P_O, T_O)$.

Ce théorème est aussi valable dans le cas où les $(Z_i)_{i\in U}$ ne sont pas indépendantes, mais seulement échangeables (i.e : la distribution jointe de $(Z_{\pi(1)}, \ldots, Z_{\pi(N)})$ est la même pour toute permutation π de $\{1 \ldots N\}$). Si ρ est le même coefficient de corrélation entre les $(Z_i)_{i\in U}$, $(-\frac{1}{N-1} \leq \rho < 1)$, alors $R(P_O, T_O) = \frac{\sigma^2}{n}(1-\rho)(1 - \frac{n}{N}\Delta)$. Si les $(Z_i)_{i\in U}$ ne sont pas échangeables, alors (P_O, T_O) est optimale parmi toutes les stratégies (P,T) telles que définies dans TH10, sauf qu'en plus T est linéaire ([2], p.97).

b) Pour illustrer l'importance d'un tel résultat, nous allons supposer que l'on est sous des conditions particulières où un résultat très connu est valable : l'allocation de Neyman. Nous constatons qu'alors, la stratégie (P_O, T_O) est aussi bonne sinon meilleure que la stratégie classique. Le théorème 10 établit donc un résultat valable sous des conditions bien plus générales que celles de l'allocation de Neyman.

Supposons que nous nous sommes contentés, dans l'analyse de la variance, de considérer les variables populations non transformées. Supposons que nous avons considéré un facteur F, non nécessairement équilibré, qui divise la population en r strates $\tau_j (j=1 \ldots r)$ d'effectifs respectifs L_j. Si, comme nous avons vu dans (A-II-7) la loi de répartition est la lognormale, nous avons beaucoup de chances d'aboutir à un résultat du genre :

$$\forall j=1\ldots r,\ \forall i\in \tau j \qquad E_M(Y_i) = \mu_j$$
$$V_M(Y_i) = \sigma_j^2$$

les $(Y_i)_{i\in U}$ indépendantes, et σ_j proportionnel à μ_j, soit $\mu_j = C \cdot \sigma_j$ où C est une constante.

Posons $$a_j = \frac{N\,\sigma_j}{\sum_{j'=1\ldots r} L_{j'} \cdot \sigma_{j'}}$$

Nous avons : $\frac{1}{N} \sum_{j=1\ldots r} \sum_{i\in\tau} a_j = \frac{1}{N} \sum_{j=1\ldots r} L_j\, a_j = 1$

D'autre part : $E_M(Y_i) = \mu_j = C \cdot \sigma_j = C \cdot a_j \cdot (\frac{1}{N} \sum_{j'=1\ldots r} L_{j'} \cdot \sigma_{j'})$

$= a_j \cdot \mu$

Nous sommes bien dans le modèle M1, tous les b_i, $i \in U$ *étant nuls.*

Du TH 10 nous savons alors que la stragie (P_O, T_O) *est optimale parmi toutes les stragies (P,T) telles que décrites dans le TH, avec :*

- P_O *est de taille fixe n, et de probabilités d'inclusion :*

$$\alpha_i = n \, \sigma_j / \sum_{j'=1\ldots r} L_{j'} \cdot \sigma_{j'}$$

- $$T_O(D) = \frac{\sum_{j=1\ldots r} L_j \cdot \sigma_j}{N.n} \sum_{j=1\ldots r} \tau_{Yj/\sigma_j}$$

$$\text{avec} \quad \tau_{Yj} = \sum_{i \in s \cap \tau_j} Y_i$$

le risque de (P_O, T_O) *s'écrit alors :*

$$\frac{1}{N^2} \left[\frac{1}{n} \left(\sum_{j=1\ldots r} L_j \, \sigma_j \right)^2 - \sum_{j=1\ldots r} L_j \, \sigma_j^2 \right]$$

Dans la méthode classique, le plan de sondage stratifié P1 s'effectue en prélevant un échantillon s_j *de taille* ν_j *de chaque strate* τ_j, *les échantillons étant tirés de façon indépendante d'une strate à l'autre, et les prélèvements équiprobables dans chacune d'elle. L'estimateur considéré est alors :*

$$T_1 = \sum_{j=1\ldots r} \frac{L_j \, \tau_{Yj}}{N \, \nu_j}$$

sous le modèle M1, nous avons :

$$R(P_1, T_1) = \frac{1}{N^2} \sum_{j=1\ldots r} L_j^2 \left(\frac{1}{\nu_j} - \frac{1}{N_j} \right) \sigma_j^2$$

sous la contrainte $\sum_{j=1\ldots r} \nu_j = n$, *ce risque est minimum pour :*

$$\nu_j = \frac{n \cdot L_j \, \sigma_j}{\sum_{j'=1\ldots r} L_{j'} \, \sigma_{j'}} \quad \text{(allocation de Neyman)}$$

Si nous supposons tous les ν_j *des entiers, alors :*

$$R(P_O, T_O) = R(P_1, T_1)$$

En général, les ν_j sont arrondis aux entiers les plus proches et alors $R(P_1, T_1) \geq R(P_0, T_0)$. (voir également [2], p.100).

c) Dans ce modèle, nous retrouvons donc le même estimateur que dans le modèle à population fixe (IV). Si nous pouvons supposer qu'il existe des ceofficients (a_{k_i} , b_{k_i}) pour tout rameau i de tout âge k tels que :

$$E_M\, Y_{k_i} = a_{k_i}\, \mu + b_{k_i}$$

$$V_M\, Y_{k_i} = a^2_{k_i}\, \sigma^2$$

(soit que , en appliquant une transformation linéaire, toutes les populations ont même moyenne et même variance), et si les autres hypothèses sont respectées, le TH10 reste valable pour la population totale et donc pour l'estimation de la moyenne par arbre sur tout le verger. Cette hypothèse n'est pas absurde pour nos données. L'analyse de variance montre, qu'en général, nous avons la même classification des effets moyens de la Rangée x Niveau x Face. (voir notamment G18, G19, G20 dans A-II).

B-V-3) Stratégies ξ-sans biais - modèle M1

La recherche des stratégies optimales (P,T) telles que T soit ξ-sans biais est basée sur la minimisation de :

$$E_M \; E_p(T - \bar{Y})^2 = R(P,T)$$

Comme nous l'avons mentionné avant, le plan de sondage étant non informatif, les 2 opérateurs E_M et E_p sont commutatifs.

Si T ne subit aucune contrainte par rapport au plan de sondage, nous pouvons alors suivre la démarche suivante :

- d'abord déterminer T en minimisant $E_M(T - \bar{Y})^2$, parmi les estimateurs tels que $E_M(T - \bar{Y}) = 0$, conditionnellement au plan de sondage P.

- Déterminer P en minimisant $E_p \; E_M(T - \bar{Y})^2$ quand seul P varie.

Notons Y_S (resp. $Y_{\bar{S}}$) le vecteur colonne de composante Y_i , $i \in S$ (resp. Y_i , $i \in \bar{S}$) (voir II-1-b).

Soit donc P un plan de sondage vérifiant les hypothèses de II, notamment qu'il est non informatif et de taille fixe n. En fait, de nombreux résultats dans la suite sont valables aussi dans le cas où la taille est aléatoire.

Nous avons donc :

$$\bar{Y} = \frac{n}{N} \; \bar{Y}_S + (1 - \frac{n}{N}) \; \bar{Y}_{\bar{S}}$$

Similairement, tout estimateur T de $\bar{Y}$ peut s'écrire sous la forme :

$$T(D/S) \; = \; \frac{n}{N} \bar{Y}_S + \frac{N - n}{N} \; U \; (D/S)$$

avec $U(D/S)$ un estimateur de $\bar{Y}_{\bar{S}}$.

Un estimateur T de $\bar{Y}$ est alors ξ-sans biais si et seulement si U est un estimateur ξ-sans biais de $\bar{Y}_{\bar{S}}$, et nous avons :

$$E_p\ E_M(T-\bar{Y})^2 = E_p\ E_M[(1-\frac{n}{N})^2(U-\bar{Y}_{\bar{S}})^2] = (1-\frac{n}{N})^2\ E_p\ E_{M_S}\ E_{M_{\bar{S}}}(U-\bar{Y}_{\bar{S}})^2$$

où E_{M_S} est l'espérance par rapport à ξ_S et $E_{M_{\bar{S}}}$ par rapport à $\xi_{\bar{S}}$ (voir III-3).

Si ξ est complétement déterminée, alors le risque de T atteint son minimum pour :

$$U(D/S) = E_{M_{\bar{S}}}(\bar{Y}_{\bar{S}}) = \int \bar{y}_{\bar{S}}\ d\ \xi_{\bar{S}}$$

Ceci est particulièrement important dans l'estimation bayésienne où ξ est supposée connue conditionnellement à un vecteur paramètre Θ, qui est déterminé à partir des observations par l'utilisation d'un second modèle sur les paramètres.

Dans notre cas, nous écrirons :

$$V_M(T) = E_M(T-\bar{Y})^2 = (1-\frac{n}{N})^2\ E_M[(U-E_M U)^2 + (\bar{Y}_{\bar{S}} - E_M\ \bar{Y}_{\bar{S}}) - 2(U-E_M U)(\bar{Y}_{\bar{S}} - E_M \bar{Y}_{\bar{S}})]$$

Comme U ne dépend de Y qu'à travers les Y_i tels que $i\in S$, et que $\bar{Y}_{\bar{S}}$ ne dépend de Y qu'à travers Y_i tels que $i\notin S$, et comme dans nos trois modèles nous supposons les Y_i indépendantes alors :

$$E_M(T-\bar{Y})^2 = (1-\frac{n}{N})^2\ (V_M(U) + V_M(\bar{Y}_{\bar{S}}))$$

Minimiser $V_M(T)$ revient donc à choisir U tel que $V_M(U)$ soit minimal (T et U étant ξ-sans biais).

b) Dans les 2 modèles M_1 et M_2, si U est ξ-sans biais de $\bar{Y}_{\bar{S}}$, alors nécessairement :

$$E_M\ U = E_M\ \bar{Y}_{\bar{S}} = \mu\ \bar{a}_{\bar{S}} + \bar{b}_{\bar{S}}$$

Conditionnellement à S, la seule inconnue est μ. Pour minimiser $V_M(U)$ il suffit alors de déterminer le meilleur estimateur U_o ξ-sans biais de μ, et de prendre $U = U_o\ \bar{a}_{\bar{S}} + \bar{b}_{\bar{S}}$.

Les coefficients a_i et b_i, $i\in U$, étant connus, il est équivalent de considérer U_o fonction des Y_i, $i\in S$, ou des Z_i, $i\in S$. La recherche peut-être restreinte seulement aux fonctions symétriques des $\{Z_i, i\in S\}$. En effet, soit

un estimateur non symétrique de μ et de variance $V_M(\tau) = S^2$ considérons les estimateurs $\tau_\pi = \tau(Z_{\pi(1)} \cdots Z_{\pi(n)})$ où π est une permutation quelconque de $\{i_1 \cdots i_n\}$. Alors τ_π a la même distribution que τ puisque tous les Z_i sont indépendantes, identiquement distribuées.

Un estimateur symétrique τ_o de μ est alors :

$$\tau_o = \sum_\pi \frac{\tau_\pi}{n!}$$

τ_o est ξ-sans biais de μ et

$$V_M(\tau_o) = \frac{1}{(n!)^2} \left[\sum_\pi V_M(\tau_\pi) + \sum_{\pi-\pi'} [V_M(\tau_\pi)\ V_M(\tau_{\pi'})]^{1/2} \cdot \rho(\tau_\pi, \tau_{\pi'})\right]$$

$$< \frac{n!S^2 + n!(n! - 1)\ S^2}{(n!)^2} = S^2 = V_M(\tau)$$

avec $\rho(\tau_\pi, \tau_{\pi'})$ la corrélation entre τ_π et $\tau_{\pi'}$. On a égalité si et seulement si pour toutes les permutations π et π', $\rho(\tau_\pi, \tau_{\pi'}) = 1$, ou encore que τ est symétrique.

Or $\bar{Z}_S$ est un estimateur ξ-sans biais de μ, fonction symétrique de $Z_1 \cdots Z_n$. L'unicité de $\bar{Z}_S$ vérifiant ces conditions découle d'un résultat de FRASER ([2], p.101), établissant la complétude de la famille de distribution de $\{Z_1 \cdots Z_n\}$ ($Z_1 \cdots Z_n$ indépendantes, identiquement distribuées, de distribution absolument continue).

Notons alors :

$$T^*(D) = \frac{n}{N} \bar{Y}_S + \left(1 - \frac{n}{N}\right) [\bar{Z}_S \bar{a}_{\bar{S}} + \bar{b}_{\bar{S}}]$$

Nous pouvons donc énoncer :

<u>C25</u> : Si les conditions du modèle M1 sont vérifiées, alors pour tout estimateur T ξ-sans biais de $\bar{Y}$:

$$R(P, T) \geq R(P, T^*) = \sigma^2 \left[\frac{1}{n} + \frac{\Delta}{N} - \frac{2 E_P \bar{a}_S}{N} - E_P \sum_{i \in s} \left(\frac{a_i - \bar{a}_S}{N}\right)^2\right]$$

avec $\Delta = \sum_{i=1}^{N} a_i^2/N$

Deux résultats similaires ont été démontrés par CASSEL-SARNDALL-WRETTMANN desquels on s'est largement inspiré. Le premier ([2], p.112) établit l'optimalité de T^* parmi tous les estimateurs linéaires ξ-sans biais de $\bar{Y}$ dans le cas où les Z_i ont les mêmes moments d'ordre 1 et 2. Si ρ est le coefficient de corrélation entre Z_i et Z_j, $i \neq j \in U$, il faut alors multiplier l'expression de $R(P, T^*)$ par $(1-\rho)$. Le second résultat ([2] , p.116) établit le même résultat que C25 dans le cas où $a_i = 1$, $b_i = 0$, $\forall i \in U$ (alors $T^* = \bar{Y}_S$).

Pour déterminer la stratégie optimale, il suffit alors de déterminer le plan de sondage en minimisant $R(P, T^*)$ quand P varie. Soit donc à choisir P tel que $E_p[\, 2\, N\, \bar{a}_S + \sum_{i \in s} (a_i - \bar{a}_S)^2\,]$ soit maximale ; ou encore à prélever préférentiellement les échantillons tels que la moyenne des a_i ainsi que leur variance dans l'échantillon soit maximale (en accordant plus d'importance à la maximisation de $\bar{a}_S$ tant que la taille de l'échantillon est faible). Si P^* désigne le plan de sondage tel que :

$$\forall s \in S \quad P^*(S = s) = 0$$

si cette expression n'est pas maximale, tous les autres échantillons étant équiprobables, alors la stratégie (P^*, T^*) est optimale parmi toutes les stratégies ξ-sans biais de $\bar{Y}$ et sous les conditions du modèle M1.

Ce résultat est surprenant vis à vis des principes de la randomisation, surtout qu'en général, la variance par rapport au plan de sondage ne peut être calculée (certaines probabilités d'inclusion d'ordre 2 nulles : TH4). Il suppose notamment la fiabilité du modèle M1 : nous sommes sûr de l'existence des strates (aggrégats - concentrations...) ainsi que du choix des paramètres a_i, $i \in U$.

Ce point sera discuté plus dans le modèle M3. Nous proposons cependant une alternative qui a l'avantage de faire toujours intervenir la randomisation. Nous avons :

$$A = n\, E_p[2\, N\, \bar{a}_S + \sum_{i \in S} (a_i - \bar{a}_S)^2]$$

$$= 2N \sum_{i=1}^{N} a_i\, \alpha_i + (n-1) \sum_{i=1}^{N} \alpha_i\, a_i^2 - \sum_{i \neq j \in U} \alpha_{ij}\, a_i\, a_j$$

où $\{\alpha_i, \alpha_{ij}\ ;\ i \neq j \in U\}$ sont les probabilités d'inclusion de P. Si, comme nous supposons, le plan de sondage est à taille fixe, alors en remplaçant α_i par

son expression en fonction des α_{ij} , $j \neq i$ (TH1) et en remarquant la symétrie de l'expression obtenue par rapport aux indices i et j, nous arrivons finalement à :

$$A = \sum_{i<j} \alpha_{ij} \left[\frac{2N}{n-1} (a_i + a_j) + (a_i - a_j)^2\right]$$

Le choix de la stratégie ξ-sans biais minimisant le risque revient donc à choisir un plan de sondage tel que ses probabilités d'inclusion maximisent l'expression A, sous des contraintes données par le TH1, et des contraintes supplémentaires telles que la non nullité des α_{ij}, de façon à pouvoir calculer la variance par rapport au plan de sondage.

C26 : *Sous le modèle M1, la stratégie optimale parmi toutes les stratégies ξ-sans biais de $\bar{Y}$ est (P^*, T^*), où P^* est tel que ses probabilités d'inclusion d'ordre 2 vérifient les contraintes du TH1 (II, 1), soient non nulles et maximisent :*

$$\sum_{i<j \in U} \alpha_{ij} \left[\frac{2N}{n-1} (a_i + a_j) + (a_i - a_j)^2\right]$$

4) Stratégie ξ-sans biais - Modèle M2

Le modèle M2 apporte plus d'information et devrait nous permettre de trouver de meilleurs estimateurs U_o *de* μ .

Dans le cas général, notons f la fonction densité de Z_i , $i \in U$, *et* $f_S(z_S) = \prod_{i \in S} f(z_i)$ *celle de* $\{Z_i ; i \in s\}$. *Dans ce modèle, nous supposons que f est entièrement déterminée conditionnellement aux paramétres* $\Theta \in \textcircled{H}$. *Nous pouvons toujours considérer que* μ *est le premier de ces paramètres. (En général,* $\textcircled{H} \subset R$ *ou* $\textcircled{H} \subset R^2$*). Alors sous des conditions générales (dérivabilité de f par rapport aux paramètres, et commutativité des opérateurs E et* $\frac{\delta}{\delta\Theta}$*),* $V_M U_o$ *vérifie ([4], p.494) :*

$$V_M U_o \geq \frac{1}{1 - \rho^2} \quad \frac{1}{n E_M\left\{\frac{\delta \text{ Log } f_S}{\delta \mu}\right\}^2}$$

avec $\rho = 0$ *si f ne dépend que du paramètre* μ .

Sinon,

$$\rho = \text{Cor}\left(\frac{\delta \text{ Log } f_S}{\delta \mu} , \frac{\delta \text{ Log } f_S}{\delta \tau}\right) \text{ (avec } \Theta = (\mu , \tau))$$

(Dans le cas $\Theta = \mu$, *cette inégalité équivaut à :*

$$\text{Corr}\left(U , \frac{\delta \text{ Log } f_S}{\delta \mu}\right) \leq 1)$$

Si nous avons égalité, l'estimateur U_o *est dit efficace, et est optimal parmi tous les estimateurs ξ-sans biais de* μ. *Dans ce cas, il peut toujours être défini par la méthode du maximum de vraisemblance ([4] , p.482). Si l'égalité n'est vérifiée qu'en limite quand* $n \to \infty$, *alors* U_o *est dit asymptotiquement efficace.*

De tels estimateurs n'existent pas toujours. Nous en rapportons certains :

a) Si la loi de distribution des $(Z_i , i \in U)$ *est normale,* $N(\mu , \sigma^2)$, *alors :*

. $\bar{Z}_S$ *est un estimateur efficace de* μ, *donc C26 est encore valable dans ce cas.*

. $(\bar{Z}_S , s^2)$ est exhaustive minimale complète de (μ , σ^2) avec $s^2 = \frac{1}{n-1} \sum_{i\varepsilon S} (Z_i - \bar{Z}_S)^2$. Donc en particulier que toute fonction $F(\bar{Z}_S , s^2)$ est optimale parmi tous les estimateurs ξ-sans biais de $E_M F(\bar{Z}_S , s^2)$. Ceci est important pour certaines distributions, notamment :

b) Si la loi de distribution des $(Z_i , i\varepsilon U)$ est une loi lognormale $LN(m , \Theta)$. Notons $\tau_i = Log(Z_i)$, $i\varepsilon U$. Utilisant ce qui précède, FINNEY [6] a montré que l'estimateur efficace devient :

$$\hat{\mu} = exp(\bar{\tau}_S + \frac{1}{2}\hat{\Theta}^2) \cdot g_{n-1}(\frac{1}{2}\hat{\Theta}^2)$$

avec $\bar{\tau}_S = \frac{1}{n} \sum_{i\varepsilon S} Log(Z_i) = \frac{1}{n} \sum_{i\varepsilon S} \tau_i$

$$\hat{\Theta}^2 = \frac{1}{n-1} \sum_{i\varepsilon S} (\tau_i - \bar{\tau}_S)^2$$

$$g_m(t) = \sum_{h=0}^{\infty} \frac{m^h (m+2h)}{m(m+2) \ldots (m+2h)} (\frac{m}{m+1}) \frac{t^h}{h!}$$

Cependant, l'estimateur $\bar{Z}_S$ a des bonnes propriétés tant que Θ^2, la variance des variables transformées τ_i, est assez faible (malgré la présence d'un biais, mais est consistant). En effet, l'efficacité de $\bar{Z}_S$ vaut [1] :

$$\frac{V_M \hat{\mu}}{V_m \bar{Z}_S} = \frac{n}{e^{\Theta^2} - 1} \left[e^{\Theta^2 (\frac{2}{n} - 1)} G_{n-1}(\Theta^2 \frac{n-1}{2n}) - 1 \right]$$

avec $G_m(t) = \sum_{h=0}^{\infty} \frac{M(\frac{m}{2})}{M(\frac{m}{2} + h)} \binom{m + 2h - 2}{h} t^h$

et $M(a) = \int_0^{\infty} e^{-t} t^{a-1} dt \qquad a > 0$

soit approximativement :

$$\frac{V_M \hat{\mu}}{V_m \bar{Z}_S} = \frac{\Theta^2}{e^{\Theta^2} - 1} \left[1 + \frac{\Theta^2}{2} + \frac{\Theta^4}{2n} (1 + \frac{\Theta^2}{4}) \right]$$

qui, pour $\Theta^2 \leq 0.7$ est toujours supérieure à 0.93.

Donc, tant que la variance des variables transformées est faible, nous pouvons considérer que C26 est toujours valable dans le cas où la loi de distribution est une lognormale.

Par contre, pour l'estimation de la variance, il est nécessaire de passer par les variables transformées, même pour des valeurs très petites de Θ^2 (pour $\Theta^2 = 0.1$, l'efficacité de l'estimateur sur variables non transformées est de seulement 0.79 [6]). L'estimateur employé est alors :

$$s^2 = \exp[2\,\bar{z}_S]\; g_{n-1}\left(\frac{n-2}{n-1}\;\hat{\Theta}^2\right)$$

c) D'autres résultats dans ce modèle concernent les fonctions densités f telles que :

$$f(z/\Theta_1, \Theta_2) = \frac{1}{\Theta_2}\; f_o\left(\frac{z - \Theta_1}{\Theta_2}\right) = \frac{1}{\Theta_2}\, f_o(\tau)$$

dans la suite, nous posons $\tau = \dfrac{z - \Theta_1}{\Theta_2}$

f est supposée satisfaire des conditions générales telles que l'existence des dérivées partielles d'ordre 2 par rapport aux paramètres, la validité d'interchanger l'ordre d'intégration et de dérivation et que $\lim_{\tau \to \pm\infty} \tau^2 f_o'(\tau) = 0$.

(Normale, Canchy, Laplace, logistique, valeur extrême)

Notons $z_{(1)} < \ldots < z_{(n)}$ les variables ordonnées associées aux $\{z_i\ ;\ i \in S\}$ et soit :

$$L_1(\tau) = -\, f_o'(\tau)/f_o(\tau)$$

$$L_2(\tau) = -\,[1 + \tau\, f_o'(\tau)/f_o(\tau)]$$

$$I = \begin{pmatrix} E_M L_1'(\tau) & E_M\, L_2'(\tau) \\ E_M\, \tau\, L_1'(\tau) & E_M\, \tau\, L_2'(\tau) \end{pmatrix}$$

alors la matrice d'information de Fischer s'écrit [3] : $\dfrac{1}{\Theta_2^2}$. I (en utilisant la condition que $\lim_{\tau \to \pm\infty} \tau^2 f_o'(\tau) = 0$).

Posons $F_o(\bar{\tau}) = \int_{-\infty}^{\tau} f_o(t)dt$

$$\left(J_1(F_o(\tau)) \ , \ J_2(F_o(\tau))\right) = \left(L_1'(\tau) \ , \ L_2'(\tau)\right) I^{-1}$$

$$\hat{\Theta}_1 = \frac{1}{n} \sum_{i=1}^{n} J_1\left(\frac{i}{n+1}\right) Z_{(i)}$$

$$\hat{\Theta}_2 = \frac{1}{n} \sum_{i=1}^{n} J_2\left(\frac{i}{n+1}\right) Z_{(i)}$$

Alors $(\hat{\Theta}_1 , \hat{\Theta}_2)$ est un estimateur ξ-sans biais, asymptotiquement efficace de (Θ_1 , Θ_2) $(\lim_{n\to\infty} Var(\hat{\Theta}_1 , \hat{\Theta}_2) = \Theta_2^2 \ I^{-1})$.

CHERNOFF - GASWIRTH et JOHNS [3] ont illustré ces résultats par le cas de la distribution logistique et de CANCHY. Bien sûr, dans le cas de la loi normale, nous retrouvons les estimateurs du cas a).

Dans le cas général, la moyenne μ est liée aux paramètres (Θ_1 , Θ_2) par :

$$\mu = \Theta_1 + \Theta_2 \int_{-\infty}^{\infty} \tau \ f_o(\tau) \ d\tau \ = \ C\Theta$$

où C est une application linéaire. Alors U_o dans C 26 est remplacé par $C\hat{\Theta}$ et $\lim_{n\to\infty} V_M \ U_o = \Theta_2^2 \ C \ I^{-1} \ C'$. Notons T l'estimateur correspondant obtenu. WRETTMAN ([2], p.119) a montré que le gain en efficacité quand les statistiques d'ordre sont utilisées peut-être important. Dans son exemple, pour tout plan de sondage de taille fixe n = 5, dans une population infinie ($N = \infty$), et pour une variable de loi de distribution Laplacienne :

$$R(P , T) = 0.72 \ . \ R(P , T^{*})$$

Dans le cas de la loi uniforme (l'utilisation des statistiques d'ordre est faite d'une façon complétement différente [7]), le gain en efficacité peut atteindre 39%.

Donc nous pouvons nettement améliorer notre estimation et réduire le risque en utilisant les statistiques d'ordre. U_o est remplacé dans l'expression de T^{*} par $C\hat{\Theta}$, la valeur du risque dépendant de la distribution.

B-V-5) Modèle M3

Dans le modèle M3, nous supposons qu'il existe des variables auxiliaires $X_1 \ldots X_q$ *telles que :*

$$\forall i \in U \quad E_M\ Y_i = \sum_{k=1}^{q} b_k X_{ki}$$

$$V_M\ Y_i = \sigma^2 V_i$$

où les V_i*, sont des fonctions connues de* $(X_{1i},\ldots,X_{qi})$*,* $i \in U$*. Les* Y_i *sont supposés indépendants, et* X_1 *éventuellement tel que* $\forall i \in U$, $X_{1i} = 1$.

Ce modèle s'écrit matriciellement, conditionnellement à S :

$$E_M\ Y_S = X_S\ b \qquad E_m\ Y_{\bar{S}} = X_{\bar{S}}\ b$$

$$V_M\ Y_S = \sigma^2\ \Sigma_S \qquad V_M\ Y_{\bar{S}} = \sigma^2\ \Sigma_{\bar{S}}$$

$$\text{et} \quad Cov_M\ (Y_S\ ,\ Y_{\bar{S}}) = 0$$

avec Y_S *et* $Y_{\bar{S}}$ *les vecteurs de* R^n *et* R^{N-n} *respectivement (voir 3),* X_S *et* $X_{\bar{S}}$ *les matrices n x q et (N - n) x q où chaque ligne correspond aux valeurs prises par les variables auxiliaires,* Σ_S *et* $\Sigma_{\bar{S}}$ *les matrices diagonales (n x n) et (N - n) x (N - n) respectivement, d'éléments diagonaux les* V_i*.* $b = (b_1 \ldots b_q)$ *est le vecteur des q paramètres inconnus.*

Le problème est alors de trouver U tel que :

$$E_M(U - \bar{Y}_S)^2 \quad \text{soit minimal et}$$

$$E_M(U - \bar{Y}_{\bar{S}}) = 0 \quad \text{ou encore :}$$

$$E_M(U) = E_M\ \bar{Y}_{\bar{S}} = \frac{1}{N-n} \sum_{k=1\ldots q} \sum_{i \in \bar{S}} b_k X_{ki}$$

Notons $X_{k\bar{s}} = \sum_{i \in \bar{s}} X_{ki} \ / (N - n)$

$$X_{\bar{s}} = (X_{1\bar{s}}\ ,\ \ldots\ ,\ X_{q\bar{s}})$$

Alors $E_M\ U = X_{\bar{s}}\ .\ b$ *où seul b est inconnu.*

Par le TH de GAUSS-MARKOV, le meilleur estimateur de b parmi tous les estimateurs linéaires ξ-sans biais est :

$$\hat{b} = (X_S' \Sigma_S^{-1} X_S)^{-1} (X_S' \Sigma_S^{-1} Y_S)$$

D'où le résultat :

TH 11 ([2], p.127)

Sous le modèle M3, le meilleur estimateur de $\bar{Y}$ parmi tous les estimateurs linéaires ξ-sans biais est :

$$T_{RM} = \frac{n}{N} \bar{Y}_S + (1 - \frac{n}{N}) \bar{X}_{\bar{S}} \hat{b}$$

et pour tout plan de sondage P (non informatif, de taille n)

$$R(P,T_{RM}) = \frac{\sigma^2}{N^2} E_p \{ \sum_{i \in \bar{S}} V_i + (N-n)^2 (\bar{X}_{\bar{S}} (X_S' \Sigma_S^{-1} X_S)^{-1} \bar{X}_{\bar{S}}) \}$$

a) Pour expliciter les propriétés de T_{RM}, nous nous plaçons dans le cas q = 1 (une seule variable auxiliaire) et supposons que la régression passe par l'origine. Le modèle s'écrit alors :

$$\forall i \in U \qquad E_M Y_i = b x_i$$
$$V_M Y_i = \sigma^2 \nu(x_i)$$

Y_i , $i \in U$, étant indépendantes, et la fonction $\nu : x \to \nu(x)$ est connue.

L'estimateur des moindres carrés $\hat{b}$ de b s'écrit alors :

$$\hat{b} = \frac{\sum_{i \in S} x_i Y_i / \nu(x_i)}{\sum_{i \in S} x_i^2 / \nu(x_i)}$$

T_{RM} est alors connu sous le nom de l'estimateur de BREWER-ROYALL :

$$T_{BR} = \frac{n}{N} \bar{Y}_S + (1 - \frac{n}{N}) \bar{x}_{\bar{S}} \hat{b}$$

Notons la similitude, évidente dans ce cas, avec T^{**} dans 3.b.c. Si nous posons $b_i = 0$, $a_i = x_i/\bar{x}$, et si $\nu(x_i) = x_i^2$ pour $i \in U$, alors $T_{BR} = T^{**}$.

Le choix du plan de sondage minimisant le risque est résumé par :

TH 12 : [8]

Soit P un plan de sondage quelconque - et définissons P^{*} par :

$\forall s \in S \quad P^{*}(s) = 0$ *si* $\sum_{i \in s} x_i$ *n'est pas maximale. Si* $v(x)$ *n'est pas décroissante, et si* $v(x)/x^2$ *n'est pas croissante, alors :*

$$R(P, T) \geq R(P^{*}, T_{BR})$$

pour tout estimateur T linéaire, ξ-sans biais de $\bar{Y}$.

De même que dans le cas du modèle M1, ce résultat contredit l'esprit de la randomisation. L'échantillonneur se trouverait en difficulté, si après l'analyse des résultats, il s'avère que d'autres plans de sondages auraient été mieux appropriés, puisqu'on ne peut calculer la variance selon le plan de sondage [2]. *C'est le cas par exemple où le modèle M3 s'avère faux, par exemple* $E_M Y_i = \beta x_i^m$ *à la place de* $E_M Y_i = \beta x_i$ $(m \neq 1)$. T_{BR1} *est alors biaisé et* [2]

$$E_M(T_R - \bar{Y}) = \bar{x}\, b \left(\frac{\sum_{i \in S} x_i^m}{\sum_{i \in S} x_i} - \frac{\sum_{i=1}^{N} x_i^m}{\sum_{i=1}^{N} x_i} \right)$$

qui est négligeable sous un plan randomisé simple, mais qui peut être important sous le plan P^{*}.

Ce résultat suppose donc que le modèle est fiable. Son utilisation pour nos données (loi lognormale, $v(x) \simeq x^2$*), avec la longueur comme variable auxiliaire, signifierait qu'il faut toujours prélever parmi les rameaux les plus longs. Mais le doute planant sur le rôle joué par la longueur nous pousse à lui préférer un plan randomisé simple. La définition d'une stratégie telle que celle proposée dans le modèle M1 est aussi possible dans ce cas, mais n'a pas été envisagée.*

b) Le choix de ν est en général de la forme $\nu(x) = x^g$, $0 \leq g \leq 2$. Notons T_{BRg} l'estimateur T_{BR} correspondant ($T_{BR1} = \bar{Y}_S \frac{\bar{X}}{\bar{X}_S}$: l'estimateur quotient). Dans la suite, nous désignerons par M3' le modèle M3 dans ces conditions.

Un résultat intéressant quand à la robustesse de ce modèle vis-à-vis du choix de la fonction ν est le suivant [8].

TH13 :

Si $0 \leq g_o < g_1$, alors pour tout plan de sondage P de taille fixe, et sous le modèle M3', nous avons :

- *Si $\nu(x)/x^{go}$ n'est pas croissante*

$$R(P, T_{BRgo}) \leq R(P, T_{BRg1})$$

- *Si $\nu(x)/x^{g1}$ n'est pas décroissante*

$$R(P, T_{BRgo}) \geq R(P, T_{BRg1})$$

Pour l'inégalité stricte, il suffit que :

$$\exists s \in S \text{ tel que } P(s) > 0$$

$$\exists\, i \neq j \in s \text{ tels que } x_i \neq x_j$$

Pour nos données, il est donc raisonnable de considérer $\nu(x) = x^2$ dans un modèle où la population sur les rameaux d'un âge est expliqué par la population totale sur les rameaux d'âge juste inférieur. (loi lognormale : la variance s^2 est liée à la moyenne μ par $s^2 = \mu^2(e^{\sigma^2} - 1)$ où σ^2 est la variance des variables transformées par $X \to \text{Log } X$).

c) Dans le cas particulier d'une seule variable auxiliaire, il est intéressant de se poser la question : Que choisir ? une stratégie (P_o, T_o) comme définie dans TH10 (voir 2) (les a_i sont proportionnels à x_i , $i \in U$) ou bien une stratégie (P, T_{BR}) comme cas particulier du TH11.

Un premier résultat dû à ROYALL [8] montre que si $\nu(x)/x^2$ n'est pas croissante, alors pour tout plan de sondage P de taille fixe n :

$$R(P, T_o) \geq R(P, T_{BR2})$$

CASSEL - SARNDALL - WRETTMAN ([2], p. 169) comparent les stratégies (P_O, T_O) *et* (P, T_{BR1}) *où P est tel que* T_{BR1} *soit P-sans biais (en plus du fait qu'il est ξ-sans biais sous les conditions du modèle M'3 si* $v(x) = x$*). Si la taille de la population est infinie, et si x est répartie suivant une loi gamma (r) (de fonction densité* $f_r(x) = x^{r-1} e^{-x}/\mu(r)$ *si* $x > 0$*), supposons que* $v(x) = x^g$ *soit le vrai modèle, alors :*

$$\frac{R(P, T_{BR1})}{R(P_O, T_O)} = \frac{n(g + r - 1)}{nr + g - 1}$$

(si $nr + g - 1 > 0$*).*

Pour notre cas, le résultat de ROYALL nous pousse donc à préférer la stragie (P, T_{BR})*.*

B-V-6) *Utilisation des mesures effectuées sur* U_h, $h < k$, *comme variables auxiliaires dans l'estimation de* $\bar{Y}_k$

Nous reprenons des notations et définitions des chapitres I et II

a) *Estimation de* $\bar{Y}_2$

Remarquons d'abord que d'après l'hypothèse H4, il est équivalent de considérer les totaux ou les moyennes sur une branche.

Nous supposons que $Y_{11} \ldots Y_{1N_1}$ *sont indépendantes, de même moyenne* μ_1 *et de même variance* σ_1^2. *Notons :*

$$i \in U_2 \qquad B_i = \frac{1}{m_1} \sum_{j \in W_i \, U_1} Y_{1j}$$

$m_1 = M_{12}$ *est le nombre de rameaux de 1 an portés par toute branche de 2 ans (H4).* B_i , $i \in U_2$, *sont donc toutes indépendantes de moyenne* μ_1 *et de variance* σ_1^2/m_1. *Nous supposons alors que :*

$$Y_{2i} = b_o + b_1 B_i + \varepsilon_{2i} \qquad \forall i \in U_2$$

$$\text{avec} \quad b = \begin{pmatrix} b_o \\ b_1 \end{pmatrix} \quad \text{inconnue,} \quad \varepsilon_{2i} \ (i = 1 \ldots N_2)$$

indépendants, de moyenne nulle et de variance $V(B_i)$ *où V est connue. Notons qu'à ce niveau, beaucoup d'amélioration peut être apportée en incluant le modèle M2 ou M1.*

Y_{2i} *et* B_i *ne sont connues exactement que sur l'échantillon* S_2 *prélevé (par la contrainte d'inclusion H9). Cependant, nous disposons de plus d'information sur les* B_i, *puisque certains rameaux de 1 an ont été prélevés à part. Définissons alors* S_{12} *comme la variable aléatoire : ensemble des rameaux distincts de* U_2 *non prélevés, et portant au moins un rameau de 1 an qui a été prélevé. Soit* ν_{12} *sa taille (aléatoire).*

$$S_{12} = \{ i_1 < \ldots < i_{\nu_{12}} \in U_2 \ \text{ tels que}$$

$$i_j \notin S_2 \text{ et } W_{ij} \cap S_1 \neq O \ \text{ pour } j = 1 \ldots \nu_{12} \}$$

Alors pour tout i de S_{12}, B_i sera estimé par :

$$\hat{B}_i = \frac{1}{card(W_i \cap S_1)} \sum_{j \in W_i \cap S_1} Y_{1j}$$

d'après 3 b c (modèle M1 ξ-sans biais) appliquée dans le cas où $a_j = 1$, $b_i = 0$, $\forall i \in U$ et U l'ensemble des rameaux de 1 an portés par W_i, B_i parmi les estimateurs ξ-sans biais.

Notons $\eta_i = \frac{1}{card(W_i \cap S_1)} - \frac{1}{m_1}$

alors $E_M \hat{B}_i = E_M B_i = \mu_1$

$V_M \hat{B}_i = \eta_i \sigma_1^2$

remarquons que si $i \in S_2$, alors $\eta_i = 0$ (H9). Par contre η_i n'est pas défini pour tous i de U_2 tel que $i \notin (S_{12} \cup S_2)$.

Notons X_2 la matrice $(n_2 x n_2)$ de lignes $(1\ B_i)$, pour $i \in S_2$ ($S_2 = \{i_1 < \ldots < i_{n_2}\}$ voir II-1) Σ_2 la matrice diagonale de composante $V(B_i)$ $(i \in S_2)$, Y_{S_2} le vecteur colonne de composantes Y_i, $i \in S_2$.

Posons :

$$\bar{Y}_{12} = \frac{1}{n_2 + \nu_{12}} \sum_{i \in S_2 \cup S_{12}} Y_{2i}$$

la moyenne sur tous les rameaux de 2 ans qui ont été ou sur lesquels au moins un rameau de 1 an a été prélevé et :

$$\bar{\eta} = \frac{1}{\nu_{12}} \sum_{i \in S_{12}} \eta_i$$

Conditionnellement à X_2, l'estimateur des moindres carrés de b est :

$$\hat{b} = (X_2' \Sigma_2^{-1} X_2)^{-1} X_2' \Sigma_2^{-1} Y_{S_2}$$

$$E_M(\hat{b}/X_2) = (X_2' \Sigma_2^{-1} X_2)^{-1} X_2' \Sigma_2^{-1} E(Y_{S_2}/X_2)$$

$$= b$$

donc aussi $E_M \hat{b} = b$, $\hat{b}$ est un estimateur ξ-sans biais de b. D'après le TH de GAUSS-MARKOV, $\hat{b}$ est optimal parmi tous les estimateurs ξ-sans biais de b, conditionnellement à X_2.

$$V_m\ \hat{b} = E_m(E_m[\hat{b} - b)(\hat{b} - b)'/X_2])$$

$$= E_m\left[(X_2'\ \Sigma_2^{-1}X_2)^{-1}\ X_2'\ \Sigma_2^{-1}\ E_M\ [(Y_{S_2} - X_2\ b)(Y_{S_2} - X_2\ b)'/X_2]\ \Sigma_2^{-1}\ X_2(X_2'\ \Sigma_2^{-1}X_2)^{-1}\right]$$

$$= E_m\ (X_2'\ \Sigma_2^{-1}\ X_2)^{-1}\sigma_2^2$$

Considérons alors l'estimateur de Y_{2i}, pour $i \in S_{12}$:

$$\hat{Y}_{2i} = \hat{X}_{2i}\ \hat{b} = (1\ \hat{B}_i)\ \hat{b}$$

Nous avons :

- $\hat{b}$ ne dépend que des B_i et Y_{2i}, $i \in S_2$
- $\hat{B}_i$ ne dépend que des Y_{1j} pour $j \in W_i \cap S_1$ et $i \in S_{12}$
- $S_{12} \cap S_2 = \emptyset$
- Y_{2i}, $i \in U_2$ sont indépendants
- Y_{1j}, $j \in U_1$ sont indépendants

Donc $E_m\ \hat{Y}_{2i} = E_M\ \hat{X}_{2i}\ .\ E_M\ \hat{b} = E_M\ Y_{2i} = b_o + b_1\ \mu_1$.

$$V_M\ \hat{Y}_{2i} = (1\ \mu_1)\ [\sigma_2^2\ E_m(X_2'\ \Sigma_2^{-1}\ X_2)^{-1}\ \begin{pmatrix}1\\ \mu_1\end{pmatrix}$$

$$+ \sigma_2^2\ \sigma_1^2\ \eta_i(0\ 1)\ E_M(X_2'\ \Sigma_2^{-1}X_2)^{-1}\begin{pmatrix}0\\ 1\end{pmatrix}$$

$\hat{Y}_{2i}$ est donc un estimateur ξ-sans biais de Y_{2i}. Cependant, nous ne pouvons rien dire quand à la transmission des propriétés d'optimalités de $\hat{b}$ et $\hat{B}_i$, étant donné que les deux ne sont pas exhaustives complétes.

Considérons alors :

$$U_{12} = (\nu_{12} \sum_{i\in S_{12}} \hat{B}_i)\ \hat{b}$$

$$= \sum_{i\in S_{12}} \hat{Y}_{2i} = \nu_{12}\ \hat{b}_o + \hat{b}_1 \sum_{i\in S_{12}} \hat{B}_i$$

U_{12} est un estimateur ξ-sans biais de $\sum_{i\in S_{12}} Y_{2i}$. Sa variance se déduit facilement de ce qui précède :

$$V_M\, U_{12} = \nu_{12}(1\ \mu_1)[\sigma_2^2\, E_m(X_2'\, \Sigma_2^{-1}\, X_2)^{-1} \begin{pmatrix}1\\ \mu_1\end{pmatrix}$$

$$+ \sigma_2^2\, \sigma_1^2\, \nu_{12}\, \bar{\eta}\ (0\ 1)\ E_m(X_2'\, \Sigma_2^{-1}\, X_2)^{-1} \begin{pmatrix}0\\ 1\end{pmatrix}$$

En appliquant la même démarche que dans 3), nous avons donc l'estimateur :

$$T_{12} = \frac{1}{n_2 + \nu_{12}} \left[\sum_{i \in S_2} Y_{2i} + b_o\, \nu_{12} + b_1 \sum_{i \in S_{12}} \hat{B}_i \right]$$

qui est ξ-sans biais de $\bar{Y}_{12}$ et de $\bar{Y}_2$, et :

$$E_M(T_{12} - \bar{Y}_{12})^2 = \frac{1}{(n_2 + \nu_{12})^2}\, E_M \left[U_{12} - \sum_{i \in S_{12}} Y_{2i} \right]^2$$

$$= \frac{1}{(n_2 + \nu_{12})^2} \left[V_M(U_{12}) + V_M\Big(\sum_{i \in S_{12}} Y_{2i} \Big) - 2\, Cov_M\Big(U_{12}\, ,\, \sum_{i \in S_{12}} Y_{2i} \Big) \right]$$

Or U_{12} est indépendant des Y_{2i}, $i \in S_{12}$ donc $Cov_M\ (U_{12}\, ,\, \sum_{i \in S_{12}} Y_{2i}) = 0$

D'autre part, avec le même raisonnement, nous avons :

$$V_m\ \Big(\sum_{i \in S_{12}} Y_{2i} \Big) = E_m\ [\sigma_2^2 \sum_{i \in S_{12}} V(X_i)]$$

D'où finalement :

C27 :

$$T_{12} = \frac{1}{n_2 + \nu_{12}} \left[\sum_{i \in S_2} Y_{2i} + b_o\, \nu_{12} + b_1 \sum_{i \in S_{12}} \hat{B}_i \right]$$

est un estimateur ξ-sans biais de $\bar{Y}_{12}$ et de $\bar{Y}_2$ et :

$$E_M(T_{12} - \bar{Y}_{12})^2 = \frac{\nu_{12}\, \sigma_2^2}{(n_2 + \nu_{12})^2} \left[\sum_{i \in S_{12}} E_M\, \hat{V}(X_i) + (1\ \mu_1)\, E_M(X_2'\, \Sigma_2^{-1} X_2^{-1} \begin{pmatrix}1\\ \mu_1\end{pmatrix} \right.$$

$$\left. + \sigma_1^2\, \bar{\eta}\ (0\ 1)\ E_M(X_2'\, \Sigma_2^{-1} X_2)^{-1} \begin{pmatrix}0\\ 1\end{pmatrix} \right]$$

A titre de comparaison, si nous nous plaçions dans les conditions du modèle M1 avec $a_i = 1$, $b_i = 0, \forall\ i \in U_2$, alors le meilleur estimateur de $\bar{Y}_{12}$ et de $\bar{Y}_2$, ξ-sans biais serait :

$$\bar{Y}_{S_2} = \frac{1}{n_2} \sum_{i \in S_2} Y_{2i}$$

$$\text{et} \quad E_M(\bar{Y}_{S_2} - \bar{Y}_{12})^2 = \frac{\nu_{12}}{n_2(n_2+\nu_{12})} (b_o + b_1 \mu_1)^2$$

$$+ \frac{\nu_{12}^2 \sigma_2^2}{(n_2+\nu_{12})^2} \left[\frac{1}{n_2^2} \sum_{i \in S_2} E_M V(B_i) + \frac{1}{\nu_{12}^2} \sum_{i \in S_{12}} E_M V(B_i) \right]$$

Conditionnellement aux b_i , $i \in S_2 \cup S_{12}$, *le TH11 assure que*

$$E_M(T_{12} - \bar{Y}_{12})^2 \leq E_M(\bar{Y}_{S_2} - \bar{Y}_{12})^2$$

(la population totale considérée étant $S_2 \cup S_{12}$*). Notons alors l'équivalence entre les 2 expressions dans C27 et TH11.*

Si $V(x) = V$ *une constante, alors :*

$$E_M(\bar{Y}_{S_2} - \bar{Y}_{12})^2 = \frac{\nu_{12}}{n_2} \frac{V}{n_2+\nu_{12}} [\sigma_2^2 + (b_o + b_1 \mu_1)^2]$$

$$E_M(X_2' \Sigma_2^{-1} X_2)^{-1} = V \; E_M \begin{pmatrix} n_2 & n_2 \bar{B}_2 \\ n_2 \bar{B}_2 & \sum_{i \in S_2} B_i^2 \end{pmatrix}^{-1}$$

$$\text{avec} \quad \bar{B}_2 = \frac{1}{n_2} \sum_{i \in S_2} B_i$$

Le calcul de cette espérance n'est pas aisé, la comparaison entre les 2 estimateurs a été faite par simulation.

b) Estimation de $\bar{Y}_\ell$, $\ell = 2 \ldots L$

De façon similaire au cas a), nous définissons les notations suivantes :

$$\forall \ell = 2 \ldots L \quad \forall i \in U_\ell \quad , \quad \forall k < \ell$$

$$B_k(\ell, i) = \frac{1}{M_{k\ell}} \sum_{j \in W_i \cap U_k} Y_{kj}$$

$$S_{k\ell} = \left\{ i_1 < \ldots < i_{\nu_{k\ell}} \in U_\ell \text{ tels que : } \; ij \notin S \text{ et } W_{ij} \cap S_k \neq \emptyset \text{ pour } j = 1 \ldots \nu_{k\ell} \right\}$$

De la contrainte d'inclusion H9, nous avons :

$$S_{\ell-1,\ell} \subset S_{\ell-2,\ell} \subset \ldots \subset S_{1\ell}$$

$$\nu_{\ell-1,\ell} < \nu_{\ell-2,\ell} < \ldots < \nu_{1\ell}$$

Soit $b'_\ell = (b_{0\ell} \quad b_{1\ell} \ldots b_{\ell-1,\ell})$ le vecteur paramètre des coefficients de la régression.

$Y_{\ell i}$, ainsi que $B_k(\ell,i)$, $k = 1 \ldots \ell-1$, sont connus pour $i \in S_\ell$. Soit X_ℓ la matrice $(n_\ell \times \ell)$ de lignes $(1 \; B_1(\ell,i) \ldots B_{\ell-1}(\ell,i)$, pour $i \in S_\ell$.

Nous supposons donc :

- $Y_{11} \ldots Y_1 N_1$ sont indépendantes, de même moyenne μ_1 et de même variance σ_1^2

- $\forall \ell = 2 \ldots L$, $\forall i \in U_\ell$

$$Y_{\ell i} = X_{\ell i} \cdot b_\ell + \varepsilon_{\ell i}$$

$$= b_{0\ell} + \sum_{k=1\ldots\ell-1} b_{k\ell} B_k(\ell,i) + \varepsilon_{\ell i}$$

avec $E_M \varepsilon_{\ell i} = 0 \quad \forall \ell=2 \ldots L, \quad \forall i \in U_\ell$

$$E_M \varepsilon_{ki} \cdot \varepsilon_{\ell j} = 0 \quad \forall (k,i) \neq (\ell,j)$$

$$V_M \varepsilon_{\ell i} = \sigma_\ell^2 E_M V_\ell(X_{\ell i})$$

avec V_ℓ une fonction connue.

Notons Σ_ℓ la matrice diagonale $n_\ell \times n_\ell$ de composante $V(X_{\ell i})$, $i \in S_\ell$. De même que dans a), nous prendrons l'estimateur :

$$\hat{b}_\ell = (X'_\ell \, \Sigma_\ell^{-1} \, X_\ell)^{-1} \, X'_\ell \, \Sigma_\ell^{-1} \, Y_S$$

qui est défini dès que $n_\ell > \ell$ (pour que $(X'_\ell \, \Sigma_\ell^{-1} \, X_\ell)$ soit inversible, il suffit que les vecteurs colonnes $(1 \; B_1 \ldots B_{\ell-1})$ soient linéairement indépendants. Il suffit alors qu'il existe ℓ lignes tels que le déterminant de la matrice correspondante $\ell \times \ell$ soit non nul. Dans le cas où $\ell = 2$, il suffit qu'il existe $i \neq j$, $\in S_2$ tels que $B_1(2,i) \neq B_1(2,j)$. Nous supposons que cette propriété est toujours satisfaite.

Notons $\bar{Y}_{1\ell} = \frac{1}{n_\ell + \nu_{1\ell}} \sum_{i \in S_\ell \cup S_{1\ell}} Y_{\ell i}$

Pour tout ensemble $A \subset S_{1\ell}$, notons

$$B_k(\ell, A) = \sum_{i \in A} B_k(\ell, i)$$

$B_k(\ell, A)$ est estimé de façon récursive par :

$$\hat{B}_1(\ell, A) = \sum_{i \in A} \sum_{j \in W_i \cap S_1} \frac{Y_{1j}}{card(W_i \cap S_1)}$$

$$\hat{B}_k(\ell, A) = \sum_{i \in A \cap S_k} \sum_{j \in W_i \cap S_k} \frac{Y_{kj}}{card(W_i \cap S_k)}$$

$$+ \hat{b}_{0\ell} + \sum_{h=1}^{k-1} \hat{b}_{hk} \hat{B}_h(\ell, A - S_{k\ell})$$

et pour estimer $\bar{Y}_{1\ell}$, nous proposons l'estimateur :

$$T_\ell = \frac{1}{n_\ell + \nu_{1\ell}} \left[\sum_{i \in S_\ell} Y_{\ell i} + \hat{b}_{0\ell} + \sum_{k=1}^{\ell-1} \hat{b}_{k\ell} \hat{B}_k(\ell, S_{1\ell}) \right]$$

L'idée est d'estimer la population moyenne $\bar{Y}_\ell$ sur tous les rameaux d'âge ℓ du verger par la moyenne non seulement sur l'échantillon prélevé, mais sur l'ensemble des branches d'âge ℓ sur lesquels au moins un rameau de 1 an a été prélevé. La population moyenne par arbre sur tout le verger est alors déduite par :

$$T = \sum_{k=1}^{L} M_k T_k$$

les propriétés de cet estimateur ont été déduites de la simulation, ainsi que la comparaison avec :

$$\hat{Y}_m = \sum_{k=1}^{L} M_k \bar{Y}_{Sk}$$

De même que dans a), l'expression de $E_M(T_\ell - \bar{Y}_{1\ell})^2$, conditionnellement à $\{B_k(\ell,i) ; k < \ell ; i \in S_\ell \cup S_{1\ell}\}$ se déduit de celle de TH11 (en remplaçant N par $n_\ell + \nu_{1\ell}$ et la somme sur la population totale par la somme sur $S_\ell \cup S_{1\ell}$). La comparaison avec l'estimateur du TH10 a été fait par simulation.

B I B L I O G R A P H I E

[1] D. BRADU, Y. MUNDLAK (1970)

Estimation in Lognormal linear models
J.A.S.A. Vol. 65 p.198-211

[2] C.M. CASSEL, C.E. SARNDAL, J.H. WRETMAN (1977)

Foundations of inference in survey sampling
Wiley - Interscience

[3] H. CHERNOFF, J.L. GASTWIRTH, M.V. JOHNS (1967)

Asymptotic distribution of Linear combinations of functions of order statistics with applications to estimation
Ann. Math. Stat. Vol 38, n°1, p. 52-72

[4] H. CRAMER (1954)

Mathematical methods of statistics
Princeton university press

[5] D.H. EBBELER (1973)

A note on large - sample approximations in Lognormal linear models

[6] D.J. FINNEY (1951)

On the distribution of a variate whose Logarithm is normally distributed
Supplement au J.R.S.S., 7, (1951) p.155-61

[7] E.H. LLOYD (1952)

Least-squares estimation of location and seale parameters using order statistics
Biometrika 39, p. 88-95

[8] R.M. ROYALL (1970)

On finite population sampling theory under certain linear regression models
Biometrika 57, p. 377-387

[9] *R.M. ROYALL - J. HERSON (1973)*

Robust estimation in finite populations I
J.A.S.A. 68, p. 880-889

[10] *R.M. ROYALL - J. HERSON (1978)*

Robust estimation in finite populations II
Stratification on a size variable
J.A.S.A. 68, p. 890-893

[11] *P. SINGH - A.K. SRIVASTAVA (1980)*

Sampling schemes providing unbiased regression estimators
Biometrika (67.1), p. 205-209

[12] *S.M. STIGLER (1974)*

Linear functions of order statistics with Smooth weight functions.
Ann. Stat. 1974, Vol2, n°4, p. 676-693

[13] *R.A. WIJSMAN (1973)*

On the attainment of the Cramer-Rao lower bound
Ann. Stat. (1973) 1, n°3, p. 538-542

[14] *F. YATES (1981)*

Sampling methods for censuses & surveys
C. Griffin SCLTD - 4ème édition

B-VI-SIMULATION

Le modèle simulé est très simplifié par rapport à ce qui a été décrit dans les chapitres précédents, mais fait appel aux différents résultats de tout ce qui précède, bien que ce soit dans des cas particuliers. En particulier, nous n'avons considéré que le plan de sondage dispersé au maximum comme décrit dans le cas particulier i de B-II-4-c ; et nous avons négligé l'aspect stratification ($a_i = 1$, $b_i = 0$, pour $i \in U$ dans le TH1O de B-V-3).

Dans 1), nous présentons les programmes utilisés et les différentes valeurs choisies. Ces programmes sont écrits en FORTRAN, et donnés en annexe avec des commentaires assez explicites. Les résultats sont discutés dans 2).

B-VI-1) DESCRIPTION DES PROGRAMMES

a) Génération des populations :

Nous faisons référence à A-II, B-I, B-III-2,3 et B-V-6.

. Nous avons considéré N_L branches d'âge L, chacune porte m^{L-k} rameaux d'âge k (dans l'hypothèse H4, nous supposons donc : $\forall k \quad m_k = m$)

La loi de distribution de la variable sur les rameaux de 1 an est supposée $N_n(\mu_1 1_n , \sigma_1^2 \, v(\mu_1) \, I_n)$

$$\forall i \in U_1 \quad Y_{1i} \sim N(\mu_1 , \sigma_1^2 \, v(\mu_1))$$

Les variables sur rameaux plus vieux sont supposées telles que :

$$\forall i \in U_k \quad Y_{ki} = \left(1 \quad B,(k,i) \ldots B_{k-1}(k,i)\right) \begin{pmatrix} b_{ok} \\ \vdots \\ b_{k-1,k} \end{pmatrix} + \varepsilon_{ki}$$

$$= \mu_{ki} + \varepsilon_{ki}$$

avec $E_M \, \varepsilon_{ki} = 0$

$$E_M(\varepsilon_{ki}^2 / \mu_{ki}) = \sigma_k^2 \, v(\mu_{ki})$$

tous les $(\varepsilon_{ki})_{\substack{k = 1 \ldots L \\ i \in U_k}}$ sont supposés indépendants

. Le générateur de nombres pseudo-aléatoires utilisé (FUNCTION RANF) est donné par OAKENFULL [6]. La méthode est connue sous le nom de registres à décalage bouclé (generalized feed back register) : les nombres générés ont des bonnes propriétés statistiques et une période très grande ($2^{98} - 1$) (voir également [2] et [4]).

Si $x_1 \ldots x_p$ sont P réalisations indépendantes d'une loi uniforme sur [0 , 1], alors la variable :

$$y = (x_1 + \ldots + x_p - \frac{P}{2}) \sqrt{\frac{12}{P}}$$

tend en loi vers une normale N(0 , 1) quand $P \to \infty$. La convergence étant rapide, nous avons pris p = 12 [3]. Une variable aléatoire y de loi $N(m , \sigma^2)$ sera donc générée par :

$$m + (x_1 + \ldots + x_{12} - 6) \cdot \sigma$$

. *La génération des populations sur les rameaux du verger est faite d'une façon récurente suivant le modèle décrit en a - (sous-broutine RAMIF).*

b) *Plan de sondage*

Ici, nous utilisons les résultats du chapitre B-II pour les probabilités d'inclusion, et III-4 pour la fonction coût.

Nous avons considéré le plan de sondage avec un coût total fixé (III-1-a). La fonction coût a été simplifiée en ne tenant compte que des dommages causés au verger et dans le cas d'une dispersion maximale (II-4-c cas particulier i) i.e : deux rameaux de même âge portés par le même rameau d'âge juste supérieur ne sont prélevés ensemble que si un rameau porteur l'a été (tant que la taille est faible). Quand au rapport des coûts pour différents âges des rameaux, nous avons pris (avec les notations IV-4) :

$$\forall k = 2 \ldots L \quad \eta = \frac{\eta_{k-1}}{\eta_k} = \frac{m_{k-1}\, C_{k-1}}{C_k} = \frac{2}{3}$$

La fonction coût s'écrit alors :

$$C_D(P) = \frac{M_1\, C_1}{\eta^{L-1}} \left[n_L + (N_L - n_L) \sum_{k=1 \ldots L-1} \eta^{L-k} (\beta_{k+1} - \beta_k) \right]$$

Soit encore, en tenant compte du fait que :

$$M_k = m^{L-k} \quad \text{et de} \quad C13 :$$

$$C_D(P) = \frac{C_1}{\eta^2} \left[n_1 + \frac{1}{m} \eta(1-\eta) \sum_{k \geq 2} (\frac{m}{\eta})^k \eta_k \right]$$

posons $C = C_D(P) \cdot \frac{m.\eta^2}{C_1}$

à chaque valeur du couple $(C , \frac{1}{\eta})$, le programme TAILLE donne toutes les tailles possibles $(n_1,\ldots,n_L)$ sous la contrainte $(n_L \geq L + 1)$, ainsi que les tailles extrêmes : n_1 maximale, n_1 maximale conditionnellement à n_2 maximale ect...

Le sous programme ECHANT calcule alors les probabilités d'inclusion pour chacune des tailles extrêmes et effectue le prélévement en faisant appel à RANF.

c) Estimation

Ici, nous faisons référence à V.

. Les coefficients de régression b_k, k=2...L, sont estimés de deux façons : moindres carrés (subroutine BT2) et moindres carrés pondérés (subroutine BT3). Dans ce dernier cas, la pondération est égale à $(1/v'(X_{ki}))$.

BT2 et BT3 font appel à la subroutine REGRES ([5] p.178) qui effectue les calculs par la méthode de GRAM-SCHMIDT modifiée ([7], p. 312).

. Pour chaque population U_k, k=2...L, $\bar{Y}_k$ est estimé de 4 façons :

- un premier estimateur (noté HT) est déduit du TH10 dans le cas ai = 1, bi = 0 $\forall i \in U$ (pas de stratification). Nous obtenons alors l'estimateur $\bar{Y}_{S_k}$: la moyenne sur l'échantillon S_k (calcul effectué par PROJ3).

- les 3 autres estimateurs correspondent à T_k tel que défini dans V-6-b :

$$T_k = \frac{1}{n_k + \nu_{1k}} \left[\sum_{i \in S_k} Y_{ki} + \nu_{1k} \hat{b}_{1k} + \sum_{h=1}^{k-1} \hat{b}_{hk} \hat{B}_h(k , S_{1k}) \right]$$

$$= \frac{1}{n_k + \nu_{1k}} \left[\sum_{i \in S_k} Y_{ki} + U_{1k} \right]$$

Le calcul de $\sum_{i \in S_k} Y_{ki}$ est fait dans PROJ3, de façon identique à celui de $\bar{Y}_{S_k}$

U_{1k} est calculé de façon récursive par PROJ4 en utilisant les vrais coefficients de régressions, ceux estimés par BT2 et ceux par BT3. Nous obtenons alors les estimateurs correspondants, notés respectivement : BD, MC, Pond.

Finalement, les estimateurs de Y_m sont fournis par PROJ3.

d) Supposons que le plan de sondage P est fixé (tailles fixées). Soit T un des 4 estimateurs considérés. Notons t_{ij} la valeur de T quand l'échantillon i a été prélevé de la population j, et soit y_j la vraie valeur (moyennes sur tous le verger des branches de 3 ans calculée par le sous programme MOYVAR), le risque de T est alors estimé par :

$$\sum_{i=1...NE} \sum_{j=1...NP} (t_{ij} - y_j)^2 \Big/ (NP \times NE)$$

qui, par la loi des grands nombres, si les NE échantillons et NP populations sont globalement indépendants, tend en probabilité, quand $NE \to \infty$ et $NP \to \infty$, vers la vraie valeur :

$$R(P, T) = E_P E_M (T - Y_m)^2 \; . = E_M E_P (T - Y_m)^2$$

Comme signalé avant (III - 1), nous constatons qu'il est équivalent de faire varier l'échantillon pour une population fixée, et puis de faire varier cette dernière, ou encore de changer le rôle de l'échantillon et de la population.

Nous avons choisi de se fixer un échantillon, de choisir (NP) populations, et puis de faire varier l'échantillon (NE échantillons), pour des raisons de rapidité de calcul. D'autres avantages s'y joignent notamment que l'indice j est hiérarchique à i (y_j n'est jamais exactement le même). L'organigramme général est résumé ainsi (voir page suivante).

Les résultats sont ensuite traités en utilisant la programmathèque AMANCE pour avoir graphes et histogrammes des différents estimateurs.

e) <u>Choix des valeurs des paramètres</u>

Dans la majorité des cas, les valeurs des paramètres ont été fixées à partir des traitements de données (A-III).

Dans le modèle, L est fixé à 3 ans, le nombre de branches NL à 100 et le nombre de ramifications m à 3.

Pour la fonction v, deux possibilités ont été retenues :

$$v(x) = 1 \quad \text{ou} \quad v(x) = x^2$$

TAILLE : calcul des tailles extrêmes possibles des échantillons étant donné L, N_L, m, C, G = 1/η

↓ Pour chaque combinaison de taille des échantillons

ECHANT : effectue le prélèvement des échantillons

↓

RAMIF : génère NL branches d'âge L suivant le modèle décrit dans (a) étant donné m, SC : les variances résiduelles, BT les coefficients de régression et la fonction v

↓

MOYVAR : calcule la moyenne et la variance (vraies) des populations par arbre

↓

Estimation des coefficients de régressions à partir des échantillons .

BT2
Variables centrées
↓
REGRES
$(X'X)^{-1}$ X' Y

BT3
Moindres carrés pondérés
↓
REGRES
$(X' \Sigma^{-1} X)^{-1}$ X' Σ^{-1} Y

PROJ3 : calcul de HT (th10)
↓ pour k=2...L

SSECH : calcul de S_{1k} et ν_{1k}

↓

PROJ4 : calcul de U_{1k}

PROJ1 → PROJ2
PROJ1 → SSECH → PROJ2

calcul des estimateurs de $\bar{Y}_k$, k=1...L et des estimateurs de y_m : BD, MC et Pond

Bouclage sur les populations

Bouclage sur les échantillons

Bouclage sur le choix des tailles des échan

La deuxième possibilité nous rapproche du cas de la loi Lognormale. Signalons que si $v(x)$ est égale à 1, la pondération dans BT3 est aussi égale à 1, et nous avons donc MC = Pond. Dans le cas où $v(x) = x^2$, nous avons prix la pondération égale à $[B_{k-1}(k,i)]^{-2}$, pour $k = 2...L$, $i \in S_k$.

Le choix des σ_k, $k = 1...L$ est fait conditionnellement à v. Pour $v(x) = 1$, nous avons pris $\sigma_k^2 = 1500$ $(k=1...L)$; et pour $v(x) = x^2$, $\sigma_k^2 = \frac{1}{4}\mu_k^2 (k = 1...L)$. Notons que nous avons $\sigma_1^2 \overset{\sim}{\sim} 1500$ dans les 2 cas.

Les valeurs des coefficients de régression ont été fixées à partir des résultats des régressions sur les variables non transformées dans A-II (non rapportés)

$$b_{02} = 80 \qquad b_{12} = 3$$

$$b_{03} = 120 \qquad b_{13} = 3 \qquad b_{23} = 1$$

La moyenne de la variable sur les rameaux de 1 an est fixée à $\mu_1 = 75$. (voir tableau A-II-T3).

Le nombre NE d'échantillons à prélever, ainsi que celui des populations (NP), ont été déterminés de façon à avoir assez de répétitions, mais aussi pour limiter le temps considérable de calcul. Pour NE = 6 et NP = 50, le temps moyen est d'environ 2 heures. Pour les échantillons de tailles faibles, nous avons pris NE = 20 et NP = 50.

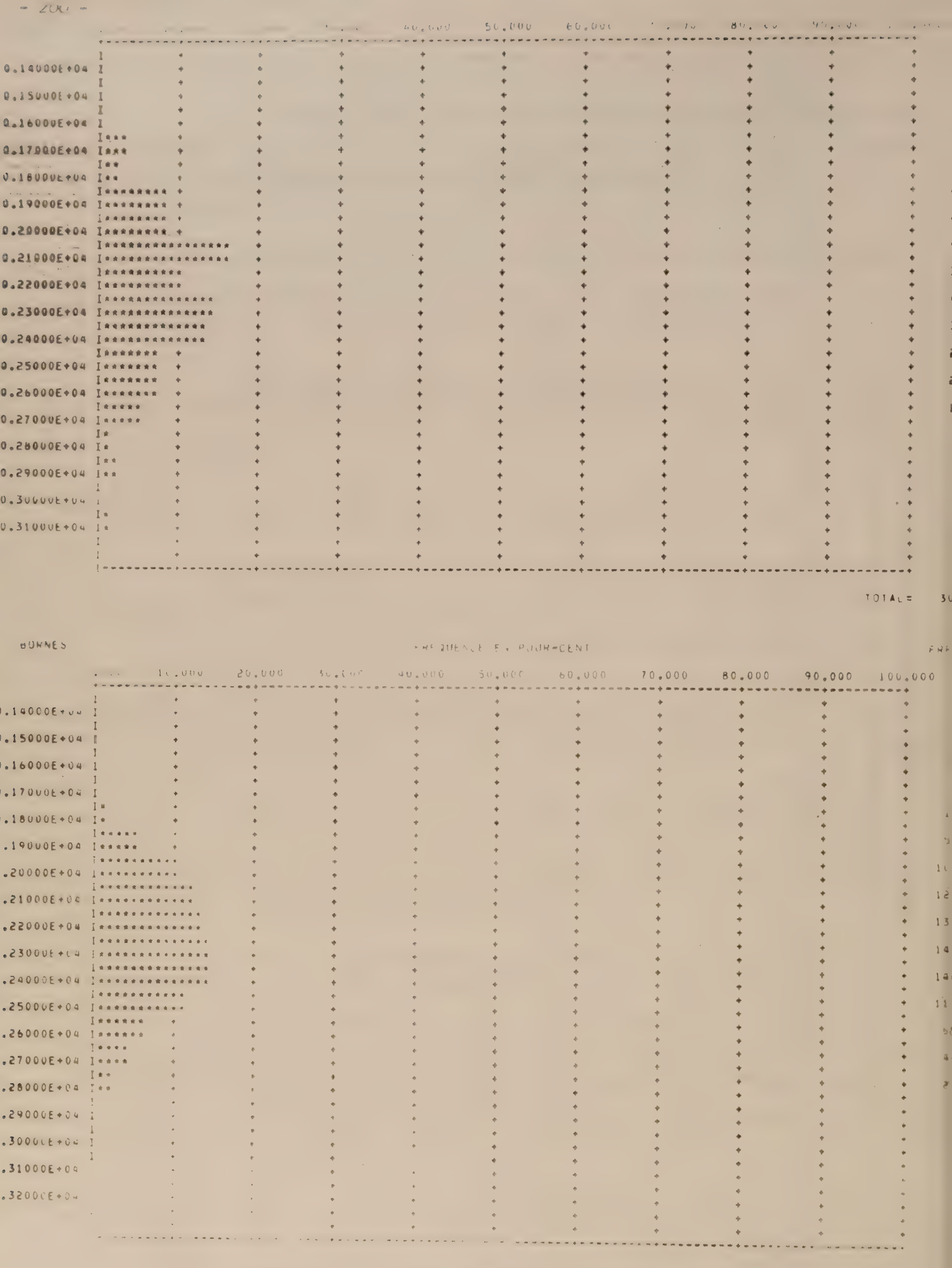
Le coût global a été fixé de façon à avoir des échantillons d'effectifs assez faibles, et tels que la taille dans le cas où on ne prélève que des branches de 3 ans soit comparable à celle donnée par les méthodes citées dans (A-I-2)(le paramètre η est fixé à 2/3).

Les tailles retenues sont présentées dans le tableau suivant (T1). Dans la suite, nous nous y référons par la numérotation indiquée colonne 3. Nous attirons l'attention sur le fait que cette numérotation est faite pour un coût donné, et est croissante avec le nombre de branches de 3 ans prélevées, puis celles de 2 ans, ce qui ne permet pas d'établir un lien direct entre deux choix de tailles ayant les mêmes numéros pour des coûts différents.

TABLEAU 1 : TAILLES RETENUES DES ECHANTILLONS

(Attention : pas de comparaison entre des numéros de tailles pour différents coûts)

Référence utilisée dans la suite	Coût global C	Taille maximale si on ne prélève que des branches de 3 ans	Nombre de branches de 3 ans n_3	Nombre de branches de 2 ans prélevées en plus ν_{23}	Nombre de rameaux de 1 an prélevés en plus ν_{12}	Nombre de branches de 3 ans sur lesquelles au moins un prélevément a été effectué $n_3 + \nu_{13}$
1	500	8	4	1	81	86
2			4	18	5	27
3			6	1	31	38
4			7	2	16	25
21	700	12	5	2	123	100
22			6	6	85	97
23			7	1	87	95
24			7	19	6	32
25			9	2	42	53
26			11	1	6	18
31	1 000	17	5	2	223	100
32			7	1	187	100
33			7	29	61	97
34			11	1	106	100
35			11	7	79	97
36			11	15	43	69
37			11	23	7	41
38			14	1	45	60
39			14	10	5	29
40			15	1	25	49



GRAPHIQUE 1 : *Histogramme de HT, cas* $v(x) = x^2$, *taille 1, avec 300 répétitions et 1000 répétitions*

2) RESULTATS

a) L'examen des histogrammes de chacun des estimateurs, dans les différents cas considérés, donne une idée sur sa loi, et indique si le nombre de répétitions est suffisant. Un exemple de cette dernière utilisation est représenté dans G1 : dans les cas des échantillons à faibles effectifs, nous avons été amené à prendre NE = 20 et NP = 50 (1000 répétitions). Ceci est dû au fait qu'alors les 3 coefficients de régression ne sont estimés qu'à partir de 4 observations, d'où une très grande variabilité de ces coefficients qui induit une grande variabilité des estimateurs associés qui dépend de ν_{13} (nombre de branches de 3 ans non prélevées et portant au moins un rameau de 1 an prélevé).

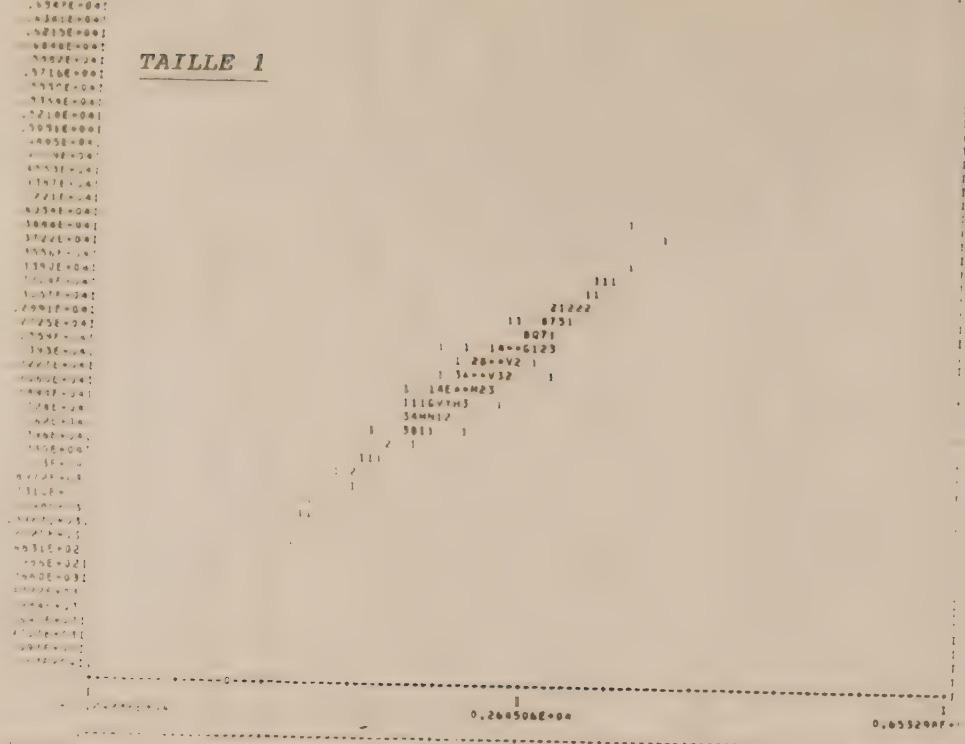
Une première constatation est que, dans tous les cas considérés ($v(x) = x^2$), les 2 estimateurs MC et Pond (déduits par l'utilisation des coefficients de régression estimés sans et avec pondération respectivement) sont équivalents (voir G2). Ceci peut être une conséquence des choix des valeurs des paramètres, et notamment des tailles.

Nous restreindrons donc l'exposé des résultats aux 3 estimateurs : HT(moyenne de l'échantillon), BD (coefficients de régression connus) et MC (coefficients de régression estimés par BT2).

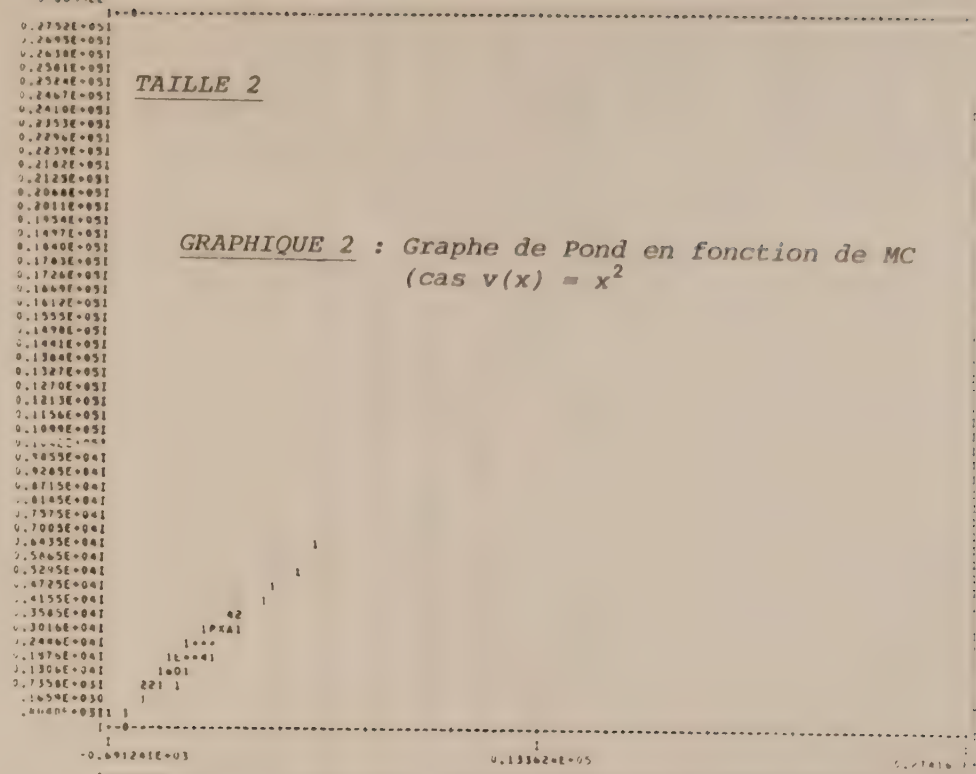
Les résultats des tests de l'hypothèse de normalité des estimateurs, basés sur les 2 coefficients de PEARSON ([1], p. 66) sont résumés dans les tableaux T2 et T3. Cette hypothèse n'est en général pas rejetée. Les cas particuliers où elle l'est correspondent à ceux où le nombre de branches de 3 ans prélevées est faible, cas où justement les risques sont particulièrement grands ; le rejet peut n'être alors qu'une conséquence du nombre de répétitions.

Le tableau T2 résume les différents résultats dans le cas $v(x) = 1$, et T3 dans le cas $v(x) = x^2$. Les valeurs des risques sont reportés dans les graphes correspondants G2 et G3.

TAILLE 1

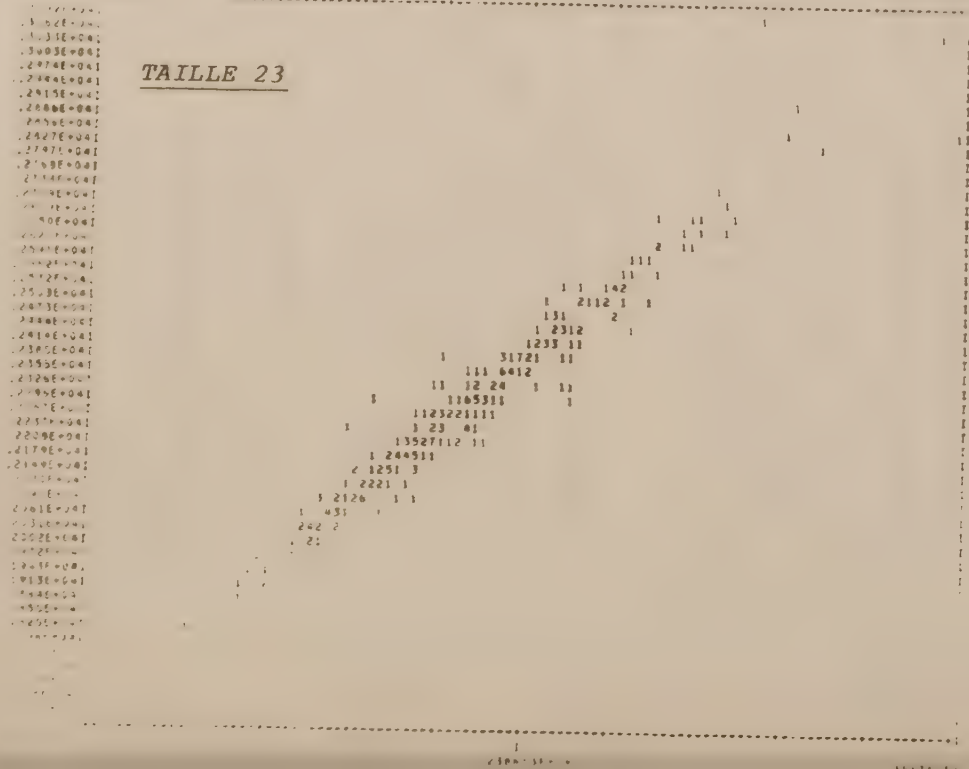


TAILLE 2

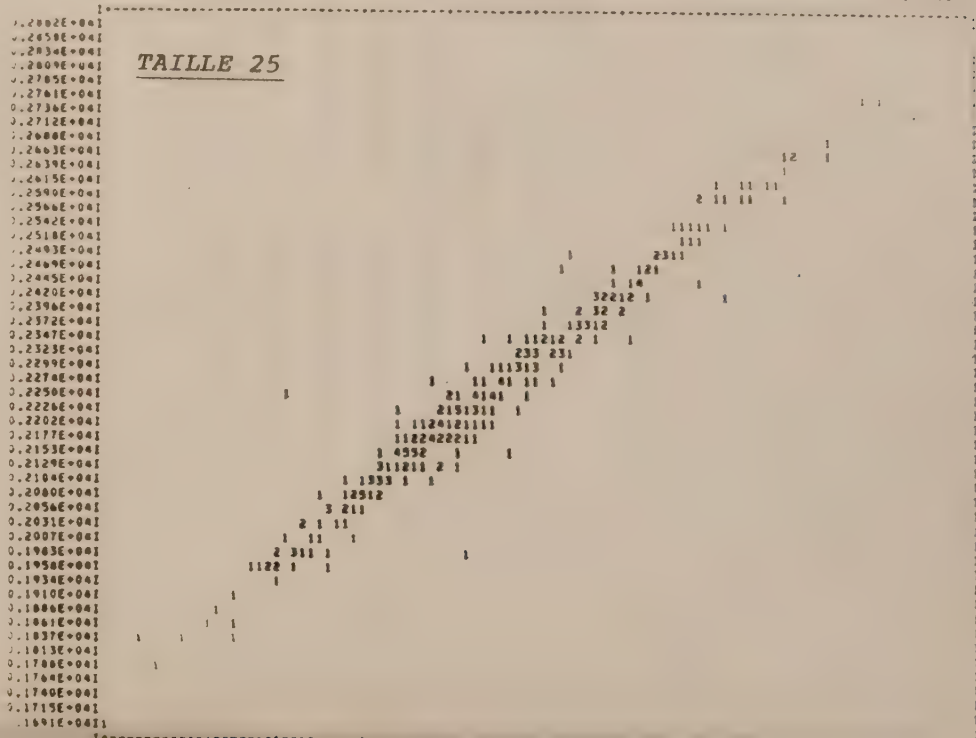


GRAPHIQUE 2 : Graphe de Pond en fonction de MC (cas $v(x) = x^2$

TAILLE 23



TAILLE 25



b) Considérons d'abord l'estimateur HT. Il est indépendant de S_{13} et S_{23} et sa variance n'est pas fonction de ν_{13}, ν_{12} et ν_{23}. Par contre, c'est une fonction croissante de :

$$\left(\frac{1}{n_3} - \frac{1}{N_L}\right) \text{ Var } Y_3 \ , \ \left(\frac{1}{n_2} - \frac{1}{m.N_L}\right) \text{ Var } Y_2 \text{ et } \left(\frac{1}{n_1} - \frac{1}{m^2 N_L}\right) \text{ Var } Y_1$$

où n_1, n_2 et n_3 sont le nombre de rameaux totaux prélevés.

Dans le cas $v(x) = 1$, nous avons :

$$\text{Var } Y_1 = \sigma^2 = 1500 \approx (38,7)^2$$

$$\text{Var } Y_2 = \left(\frac{1}{m} b_{12}^2 + 1\right) \sigma^2 \approx (81,6)^2$$

$$\text{Var } Y_3 = \frac{\sigma^2}{m^2} \left[(b_{13} + b_{12} b_{23})^2 + m\, b_{23}^2 + m^2\right] \approx (68,3)^2$$

Dans le cas $v(x) = x^2$, nous obtenons :

$$\text{Var } Y_1 = \frac{1}{4} \mu_1^2 = (37,5)^2$$

$$\text{Var } Y_2 = \frac{1}{4} \left[b_{02}^2 + \frac{5}{4m} b_{12}^2 \mu_1^2 + 2\, b_{02}\, b_{12}\, \mu_1\right] \approx (126)^2$$

$$\text{Var } Y_3 = \frac{1}{4} E \left[b_{03} + b_{13} \frac{1}{m^2} \sum_{j \in W_i \cap U_1} Y_{1j} + b_{23} \frac{1}{m} \sum_{j \in W_i \cap U_2} Y_{2j}\right]^2$$

$$= \frac{1}{4} \left[b_{03}^2 + \frac{5}{4} \frac{1}{m^2} b_{13}^2 \mu_1^2 + 2\, b_{03}\, \mu_1 (b_{13} + b_{03}^2)\right.$$

$$+ \frac{5}{4} \frac{1}{m} b_{12}^2 b_{23}^2 \mu_1^2 + 2\, b_{02}\, b_{13}\, b_{23}\, \mu_1 + \frac{5}{4m} b_{23}^2 (b_{02} + b_{12}\, \mu_1)^2$$

$$\left. + \frac{5}{2} b_{12}\, b_{13}\, b_{23} \frac{1}{m^2} \mu_1^2\right] . \approx (216)^2$$

Dans les 2 cas, $v(x) = 1$ ou $v(x) = x^2$, la variance des populations sur les rameaux de 1 an (Var Y_1) est faible comparée à Var Y_2 et Var Y_3. Ce qui explique que R(HT) est plus sensible aux variations de n_3 et n_2 qu'à celle de n_1 (voir par exemple tailles 2 et 3 du tableau T3). Remarquons finalement que dans le cas $v(x) = 1$, c'est Var Y_2 qui est la plus élevée, alors que dans le cas $v(x) = x^2$, c'est Var Y_3.

c) Les 2 estimateurs BD et MC peuvent s'écrire :

$$BD = \frac{n_3}{n_3 + \nu_{13}} HT + \tau_1 \qquad MC = \frac{n_3}{n_3 + \nu_{13}} HT + \tau_2$$

avec τ_1 et τ_2 des fonctions des valeurs observées sur les branches de 1 an et 2 ans prélevées en plus de S_3 :

$$A_1 = \{y_{1i} ; i \in \bigcup_{j \in S_{13}} W_j \cap S_1\} \cup \{y_{2i} ; i \in \bigcup_{j \in S_{23}} W_j \cap S_2\}$$

et τ_2 fonction, en plus, des estimateurs des coefficients de régression $\hat{b}_2$ et $\hat{b}_3$, déterminés respectivement à partir de :

$$A_2 = \{y_{2i} ; i \in S_2\} \cup \{y_{1i} ; i \in \bigcup_{j \in S_2} W_j \cap S_1\}$$

et de :

$$A_3 = \{y_{3i} ; i \in S_3\} \cup \{y_{2i} ; i \in \bigcup_{j \in S_3} W_j \cap S_2\} \cup \{y_{1i} ; i \in \bigcup_{j \in S_3} W_j \cap S_1\}$$

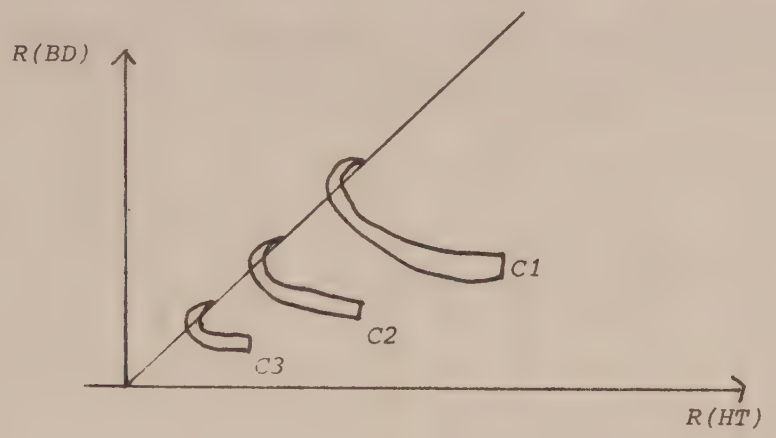
Dans la suite, nous avons pris HT comme référence dans la comparaison. Il est évident que BD est toujours meilleur que MC, puisque la seule différence entre les deux, c'est au niveau des coefficients de régression.

La corrélation entre BD et MC d'une part et HT d'autre part croit avec le nombre de branches de 3 ans prélevées. Si $\nu_{13} = 0$ (on ne prélève que des branches de 3 ans), alors BD et MC sont identiques à HT.

d) La variation du risque R(BD) (représenté par des triangles sur les graphiques) en fonction de A_1 est une fonction décroissante de $(n_3 + \nu_{13})$ (voir par exemple tailles 1 et 2, tableau T3 et 4 et 23 tableau T2). Mais plus ν_{13} est grand, plus on estime de branches de S_{13} à partir des rameaux de 1 an ou des branches de 2 ans prélevées en sus. Ainsi le rapport R(BD)/R(HT) peut dépasser la valeur 1 (BD est moins efficace que HT), c'est le cas des tailles 4 et 26 dans T2. Pour ces 2 cas, $(n_3 + \nu_{13})$ est faible (25 et 18 resp.) et ν_{23} est faible comparée à ν_{12} ($\nu_{23} = 2$, $\nu_{12} = 18$ et $\nu_{23} = 1$, $\nu_{12} = 7$ resp.). C'est à dire que les branches de S_{13} sont estimées essentiellement à partir de rameaux de 1 an ; la variation de la population sur les rameaux de 2 et 3 ans étant les plus fortes, et $(n_3 + \nu_{13})$ pas assez grand pour équilibrer, nous obtenons R(BD) plus grand que R(HT).

Dans tous les autres cas, BD est meilleure que HT et le gain en efficacité peut atteindre 90% (taille 31 du tableau T3).

Pour un coût fixé, l'évolution du risque R(BD) en fonction de R(HT), pour les différentes combinaisons de tailles possibles passe par un minimum si ν_{23} maximale conditionnellement à $(n_3 + \nu_{13})$ maximale. Cette évolution peut être représentée comme suit :



chacune des courbes correspondant aux différentes tailles possibles sous un coût fixé, avec C1 < C2 < C3 les coûts globaux. Plus le coût est grand, plus les échantillons prélevés sont grands, et plus les risques sont faibles.

Le fait qu'on n'ait pas observés de cas où BD est moins efficace que HT dans le cas $v(x) = x^2$, est dû aux valeurs des variances Var Y_1 , Var Y_2 et Var Y_3 ainsi qu'au choix des tailles.

e) La différence entre MC et BD est que MC fait intervenir des estimations des coefficients de régression. Les conséquences au niveau des valeurs des risques sont considérables (comparer les colonnes R(BD) et R(MC) des tableaux T2 et T3). Dans le tableau T2, MC est souvent meilleur que HT et le gain en efficacité peut atteindre 40 % (taille 21, tableau T2). Les graphes G3, G4 de MC en fonction BD montrent que les grandes valeurs des risques de MC sont dûes à quelques valeurs aberrantes (une dizaine sur 300 répétitions), et qu'en général on peut s'attendre à un comportement plus proche de celui de BD.

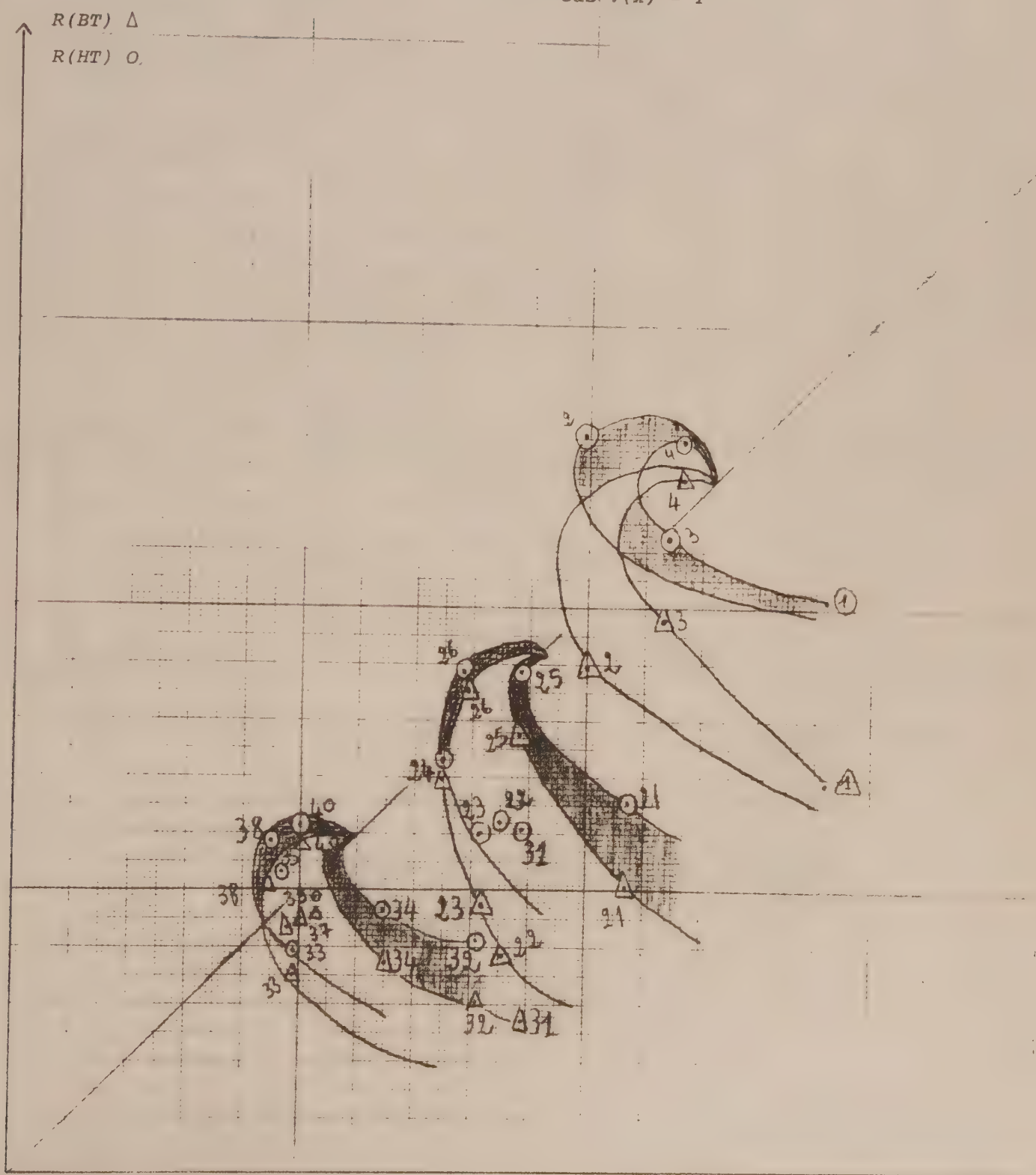
Tailles	Test de normalité			Corrélations x 100			Risque x 10^{-3}			$\frac{R(BD)}{R(HT)}$	$\frac{R(MC)}{R(HT)}$
	HT	BD	MC	(BD,HT)	(MC,HT)	(BD,MC)	R(HT)	R(BD)	R(MC)		
1			:::	48	56	86	14.2	6.9	10.1	0.49	0.72
2			:::	84	75	87	10.	8.9	13.	0.89	1.30
3				61	67	95	11.4	9.8	11.2	0.87	0.98
4			:	72	73	98	11.7	12.3	13.	1.05	1.11
21				55	62	90	10.7	5.	6.4	0.47	0.6
22				64	70	92	8.5	4.7	6.1	0.55	0.72
23				58	66	92	8.2	4.8	6.	0.58	0.73
24			:	86	86	98	7.5	7.	7.3	0.93	0.97
25				62	64	98	8.9	7.8	8.9	0.88	1.01
26				85	86	97	7.9	8.6	8.9	1.09	1.13
31	:::	:	:	39	53	82	9.6	2.7	6.1	0.28	0.64
32				49	63	89	8.1	3.	4.1	0.36	0.50
33				65	72	94	4.9	3.5	3.9	0.72	0.80
34				66	69	97	6.5	3.7	4.6	0.58	0.72
35		:	:	66	72	97	4.7	4.4	5.2	0.95	1.10
36	:			66	71	97	5.	5.5	5.5	1.09	1.09
37	:			90	90	99	5.	4.5	4.8	0.91	0.96
38				68	72	98	4.7	5	5.7	1.06	1.21
39				87	87	99	4.9	4.9	4.9	0.99	1.
40				67	72	99	5.	5.8	6.1	1.15	1.2

TABLEAU 2 : cas $v(x) = 1$ $\quad \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1500$

':::' (resp. ':') indique que l'hypothèse de normalité est rejetée au niveau 1% (resp. 5%) - l'absence d'étoiles indique que cette hypothèse n'a pas été rejetée.

GRAPHIQUE 3 : Risque de BD (coefficients de régression connus) et de MC (coefficients de régression estimés à partir des échantillons prélevés) en fonction de R(HT)

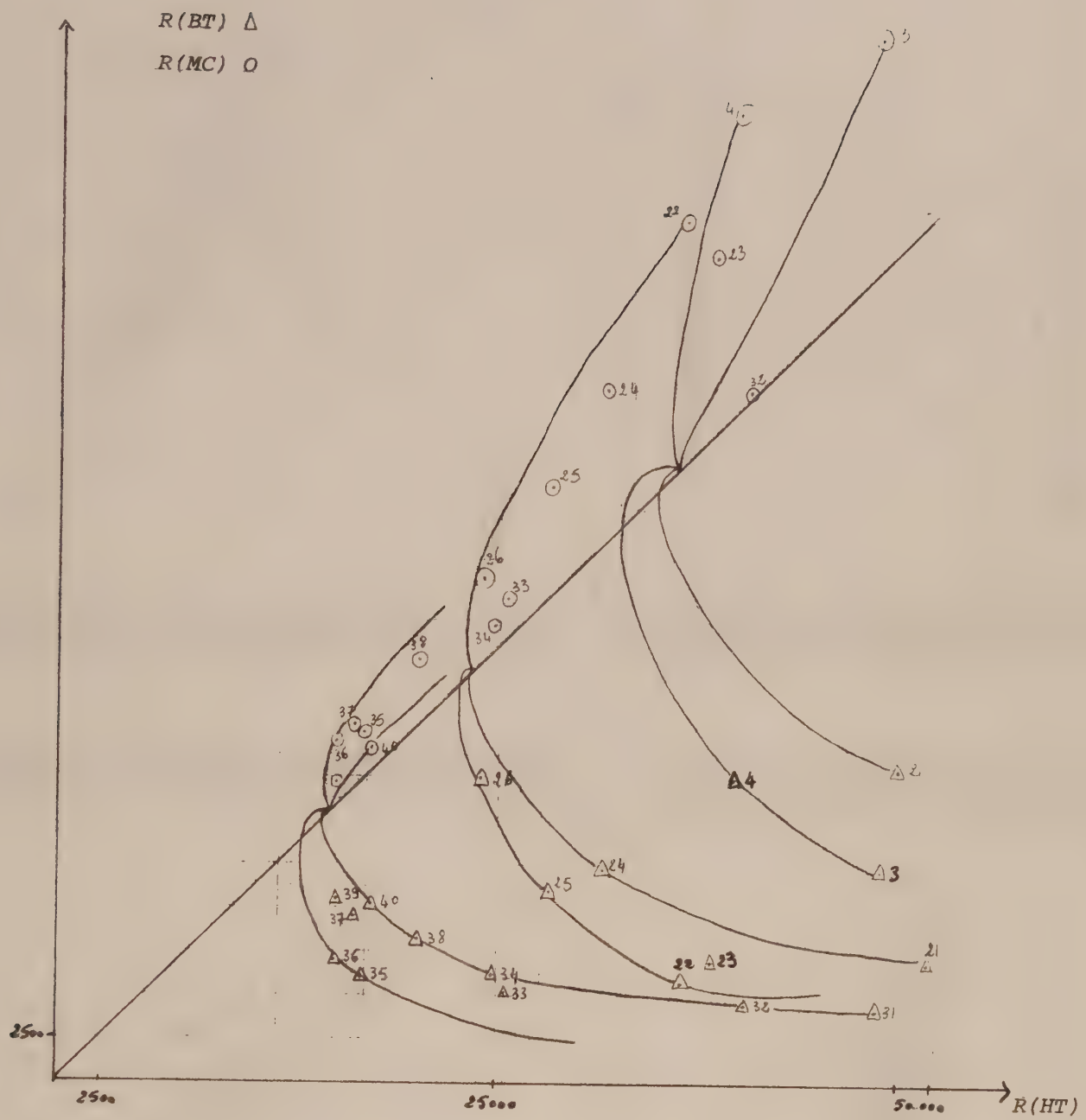
cas v(x) = 1



Tailles	Test de normalité			Corrélations x 100			Risque x 10^{-3}			$\frac{R(BD)}{R(HT)}$	$\frac{R(MC)}{R(HT)}$
	HT	BD	MC	(BD,HT)	(MC,HT)	(BD,MC)	R(HT)	R(BD)	R(MC)		
1			×××	41	56	33	60.5	9.2	151.	0.15	2.5
2	×	×	×××	72	25	22	48.2	17.9	807.	0.37	16.7
3	×		×××	62	82	66	47.1	12.2	61.	0.26	1.29
4			×××	77	81	77	38.4	17.9	55.8	0.47	1.45
21	×			54	82	57	49.4	7.2	67.7	0.15	1.37
22			×	55	79	56	35.6	5.8	49.5	0.16	1.39
23			×	55	86	54	37.2	6.95	47.4	0.19	1.27
24			×	81	84	79	31.1	14.9	39.3	0.48	1.26
25	×		×	68	81	80	28.	10.8	34.2	0.38	1.22
26				92	92	94	24.2	17.4	28.6	0.72	1.18
31				42	71	32	47.2	4.6	79.6	0.10	1.69
32				44	86	51	39.2	4.6	39.9	0.12	1.02
33		×		58	85	61	25.6	5.3	30.6	0.21	1.20
34				57	90	66	24.9	6.9	26.2	0.24	1.05
35				58	87	67	17.3	6.1	19.7	0.27	1.12
36				73	86	78	15.6	7.	18.9	0.44	1.21
37				87	91	88	16.8	9.7	20.5	0.57	1.22
38				78	88	85	20.4	8.3	24.2	0.41	1.19
39	×			91	90	92	15.8	10.5	17.4	0.67	1.11
40				80	89	89	17.7	10.4	19.1	0.58	1.08

TABLEAU 3 : cas $v(x) = x^2$ $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 0.25$

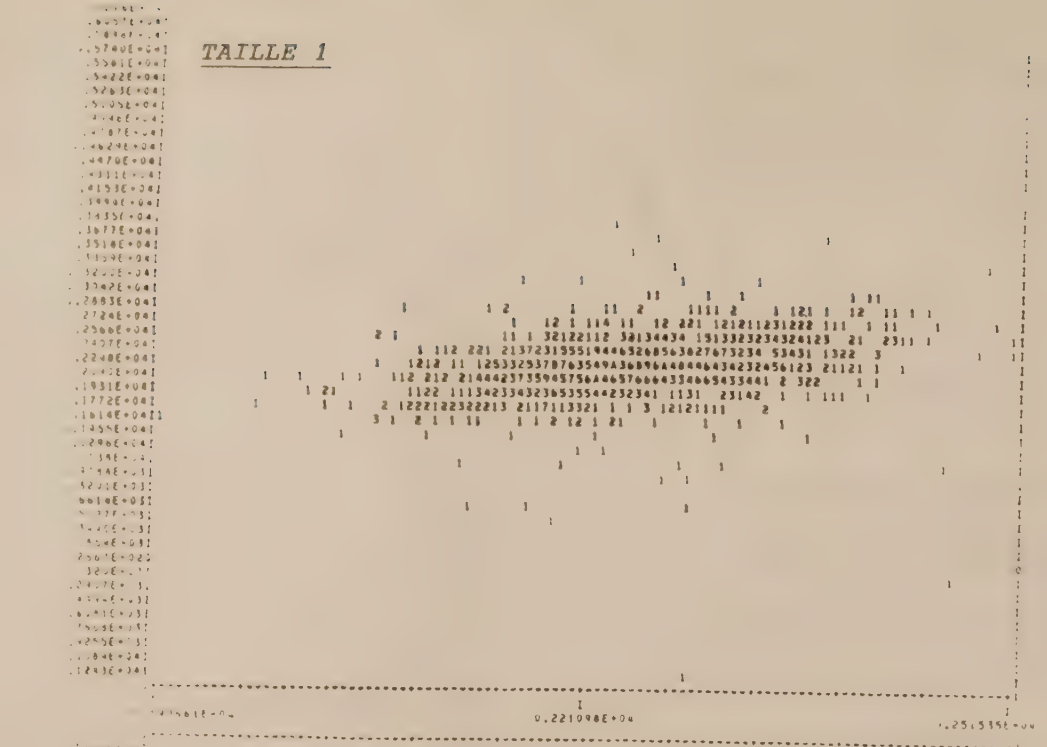
'×××' (resp. '×') indique que l'hypothèse de normalité est rejetée au niveau 1% (resp. 5%) - l'absence d'étoiles indique que cette hypothèse n'a pas été rejetée.



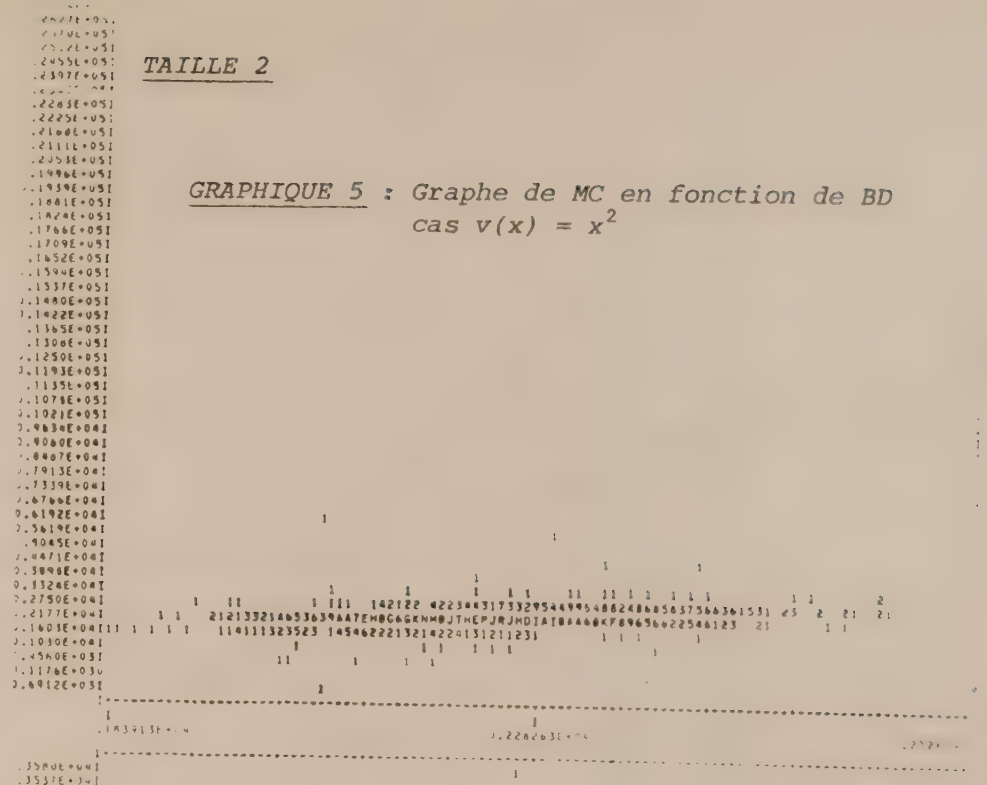
GRAPHIQUE 4 : *Risque de BD (coefficients de régression connus) et de MC (coefficients de régression estimés à partir des échantillons prélevés) en fonction de R(HT).*

cas $v(x) = x^2$

TAILLE 1

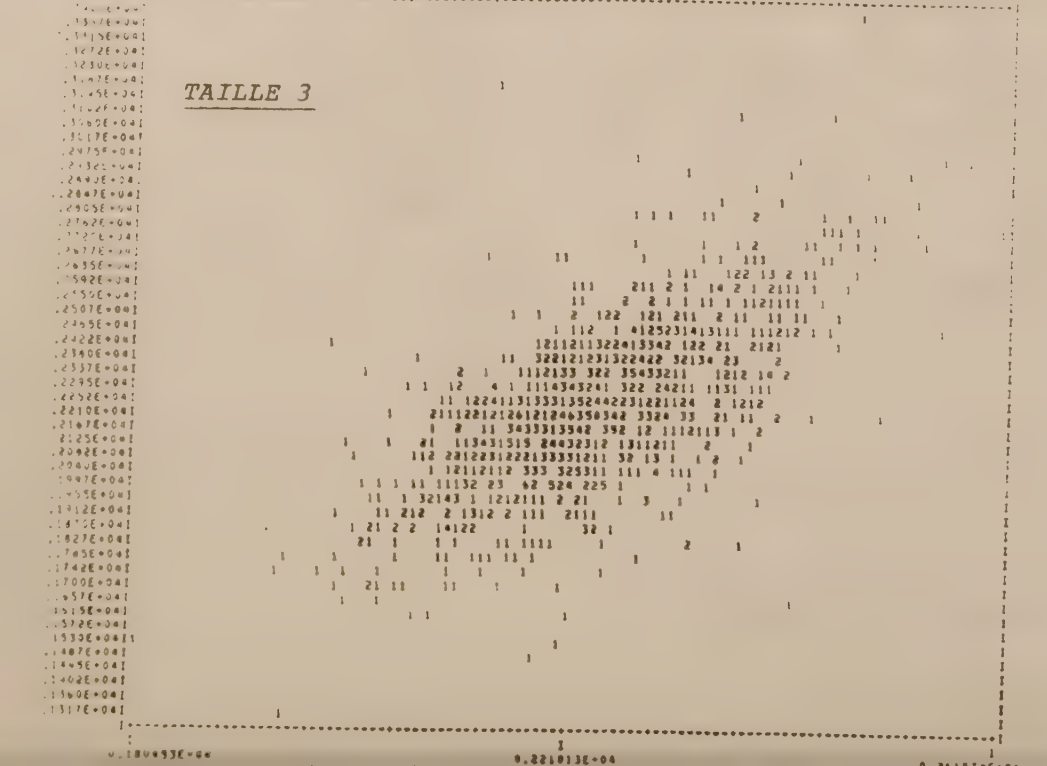


TAILLE 2

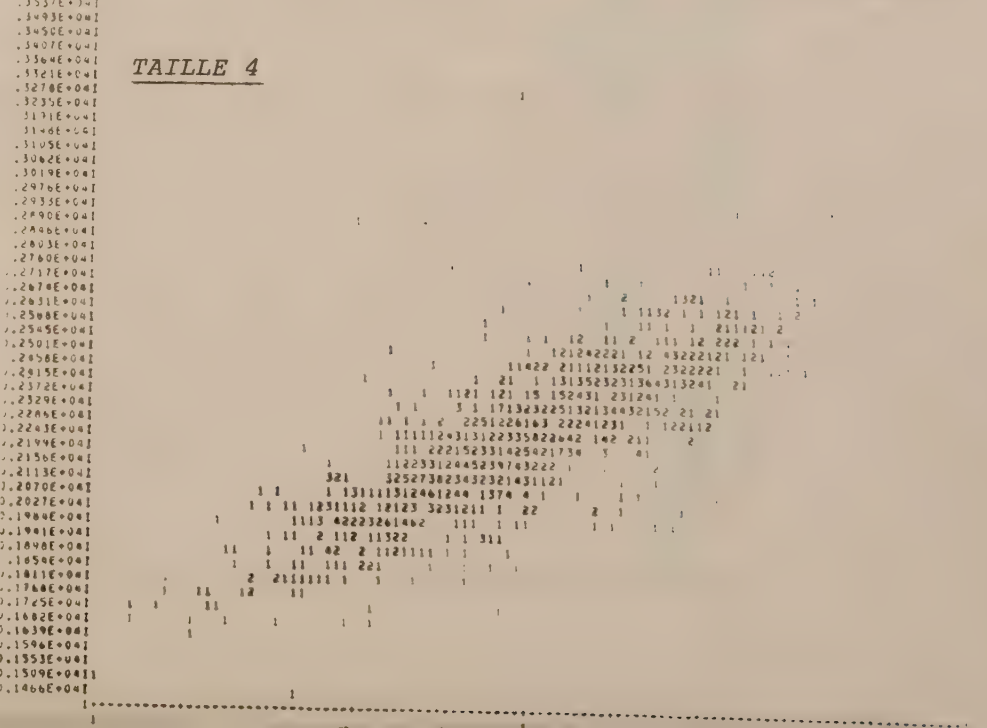


GRAPHIQUE 5 : Graphe de MC en fonction de BD
cas $v(x) = x^2$

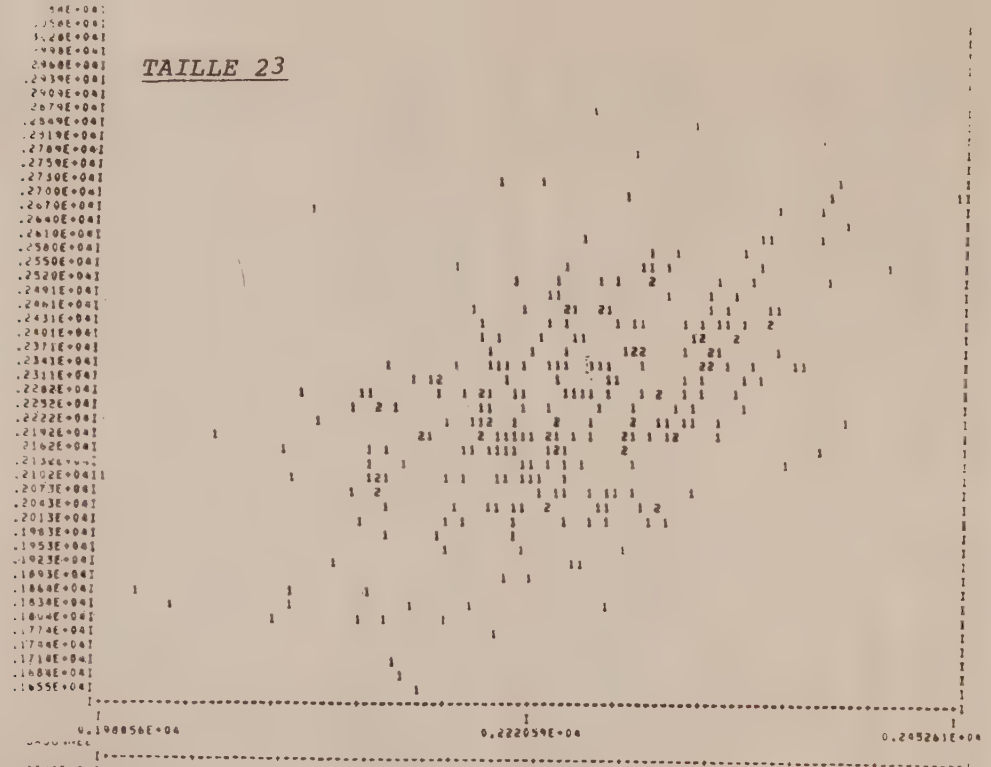
TAILLE 3



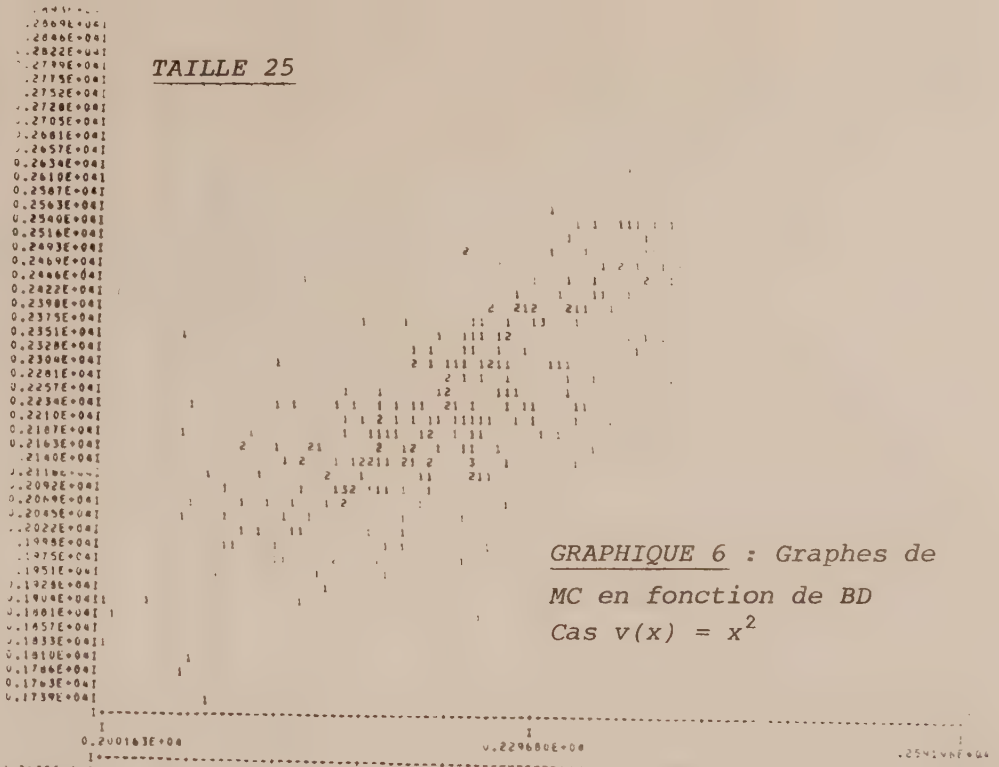
TAILLE 4



TAILLE 23

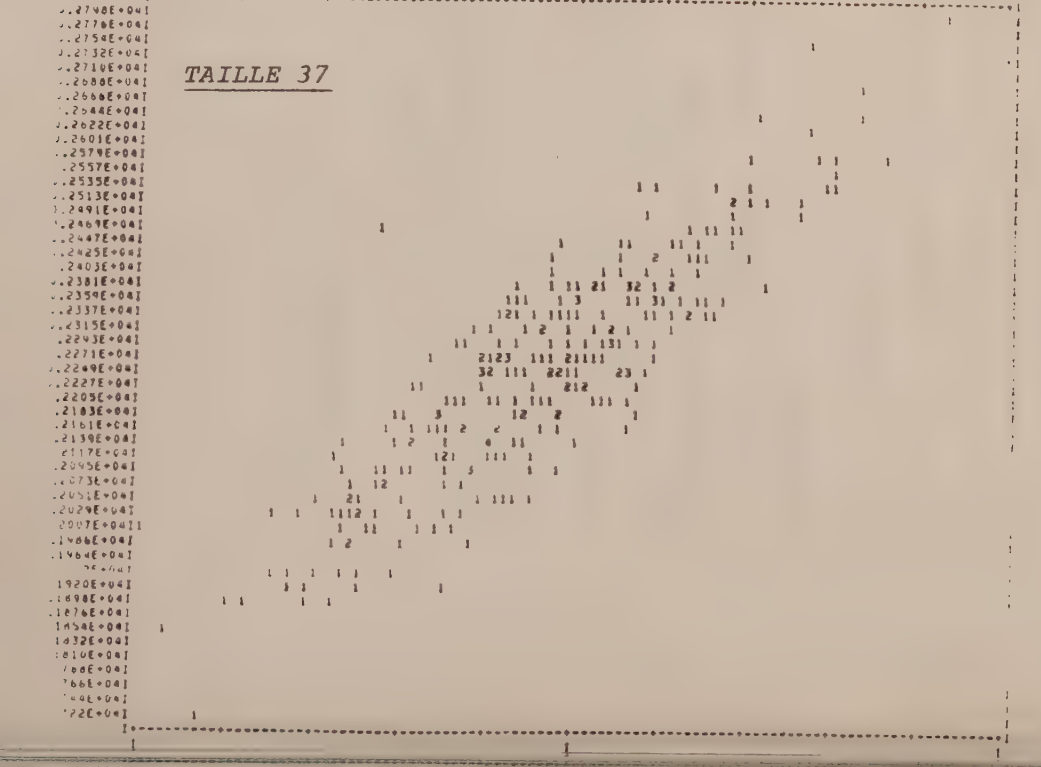


TAILLE 25

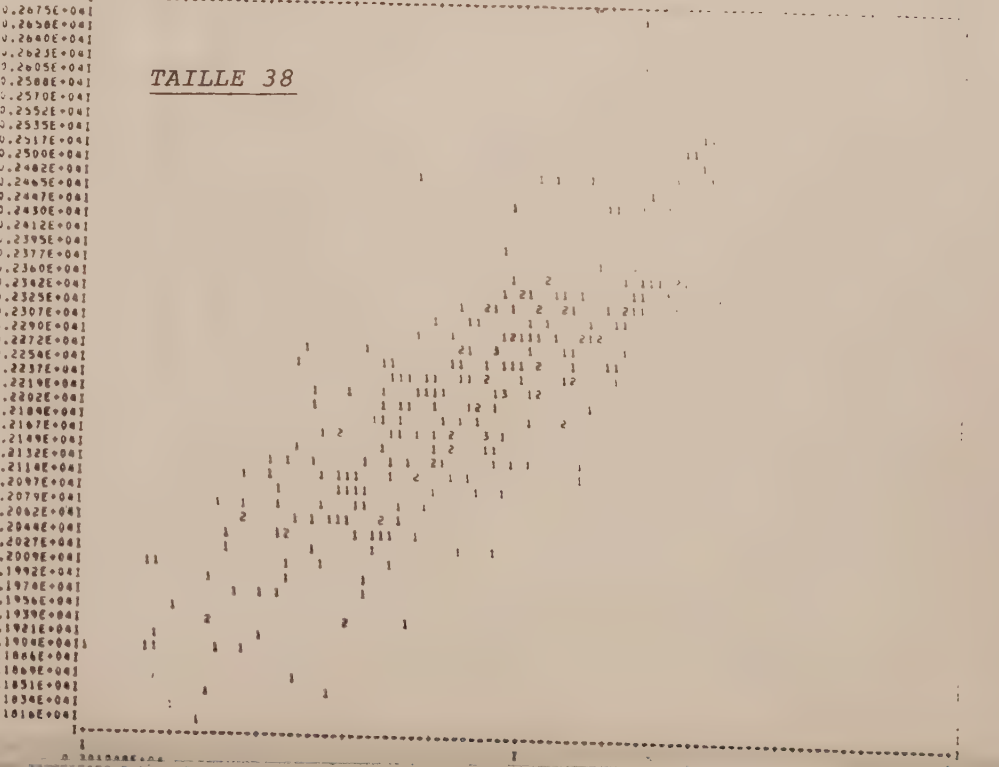


GRAPHIQUE 6 : Graphes de MC en fonction de BD
Cas $v(x) = x^2$

TAILLE 37



TAILLE 38



3 - CONCLUSION

Sous les conditions du modèle décrit en a), c'est-à-dire si on a une bonne régression (de paramètres inconnus) liant les valeurs mesurées sur un bois à celles sur ses ramifications, et avec un choix équivalent des valeurs des paramètres, nous constatons qu'il est préférable de prélever un nombre de rameaux de 3 ans "suffisant" pour assurer une estimation "correcte" des coefficients de régression et de prélever en plus des rameaux de 1 an et des branches de 2 ans de façon à ce que ν_{23} soit maximale conditionnellement à $(n_3 + \nu_{13})$ maximale. Le gain en efficacité dans le cas $v(x) = 1$ peut alors atteindre 40 %.

Dans le cas $v(x) = x^2$, il faudra faire attention au choix des coefficients de régression. La méthode utilisée dans ce cas ne correspond pas à l'esprit décrit dans B-V-4, et nous pouvons nous attendre à avoir des résultats comparables à ceux du cas $v(x) = 1$ si les coefficients de régression sont estimés à partir des variables transformées. La programmation de ce cas ne figure pas dans ce travail. Pour mener le calcul de l'estimateur de la moyenne, on fait intervenir une estimation de la variance qui peut être faite en s'aidant des techniques de JACKNIFE étant donné un échantillon.

L'estimateur utilisé jusqu'à maintenant coïncide aux 3 estimateurs HT, BD et MC quand $\nu_{13} = 0$ (seules des branches de 3 ans sont prélevées). Son risque est représenté sur les graphiques par le point d'intersection des courbes de R(BD) et R(MC) et de la première bissectrice.

B I B L I O G R A P H I E

[1] *J. BACHACOU, J.P. MASSON, C. MILLIER (1980)*

AMANCE - Manuel de la Programmathèque Statistique

[2] *W.J. KENNEDY & J.E. GENTLE (1980)*

Statistical computing
Marcel Dekker

[3] *L. LEBART, A. MORINEAU, N. TABARD (1977)*

Techniques de la description statistique
Dunod

[4] *J. LEROUDIER (1980)*

La simulation à événements discrets
Monographie de l'AFCET, Ed. Hommes et Techniques

[5] *D.R. Mc NEIL (1977)*

Interactive data analysis
J. Wiley & Sons

[6] *E. DAKENFULL (1979)*

Uniform random number generators and the spectral test
Interactive statistics, D. Mc Neil (ED)
North Holland

[7] *G.A.F. SEBER (1977)*

Linear regression analysis
J. Wiley & Sons

C O N C L U S I O N

Cherchant une méthode de sondage qui serait valable pour toute plantation, nous avons retenu du traitement de nos données (Partie A) les conclusions suivantes :

- la loi de distribution des populations sur les segments d'un âge donné peut être raccordée à une loi lognormale (ou une combinaison de lois lognormales).

- les effets de certains facteurs (Rangée, niveau) sont hautement significatifs.

- Il y a une bonne relation de régression liant la population sur un segment à celles sur ses ramifications (relation qui peut-être fonction de certains facteurs incontrôlés ...).

Nous avons aussi noté la possibilité d'avoir une bonne relation entre la population sur bourses et celle sur le bois porteur. De même, d'après la bibliographie, une voie intéressante serait de chercher à établir un lien entre le degré d'attaque et le pourcentage d'"obstacles" occupés par au moins un oeuf.

Nous avons alors considéré le problème général de sondage sur une structure ramifiée, tout en se référant aux résultats précédents pour le choix des modèles.

Ce problème a été formalisé sous l'hypothèse que tout rameau d'âge k porte le même nombre m_k de rameaux d'âge k-1 (hypothèse H4). Le plan de sondage P a été défini par ses probabilités d'inclusion d'ordre 1 et 2, en tenant compte de la liaison entre deux individus de la population (ramification) et du fait que si on coupe un rameau, alors on prélève aussi tout ce qu'il porte. Si nous nous intéressons à des branches d'âge L, il suffit de déterminer les probabilités d'inclusion pour les branches d'âge L, et des coefficients

(β_k , γ_k) pour chaque âge $k = 1...L-1$, avec β_k lié à la proportion de branches d'âge k prélevées et γ_k lié à la "dispersion" des prélèvements.

Une fonction coût global a été définie, en tenant compte de l'importance η_k de chaque âge $k = 1...L-1$, et du plan de sondage. Etant donné un coût global, et les paramètres $(\beta_k , \gamma_k , \eta_k)_{k=1...L-1}$, nous savons alors déterminer les tailles des échantillons de rameaux d'âges k, $k = 1...L$ à prélever.

Les estimateurs T de la moyenne par branche d'âge L considérés ensuite sont de la forme :

$$T = \sum_{k=1...L} M_k T_k$$

où T_k est un estimateur de la moyenne $\bar{Y}_k$ sur les rameaux d'âge k, et M_k le nombre de rameaux d'âge k porté par une branche d'âge L (supposé constant H4).

Nous avons ensuite suivi la démarche générale dans la détermination des stratégies (P_k , T_k) pour l'estimation de $\bar{Y}_k$, d'abord sous le modèle à population fixe, ensuite sous 3 modèles de superpopulations.

Sous le modèle à population fixe, la stratégie (P , T) où P est de probabilités d'inclusion $(\alpha_i , i \in U)$ tels que définies dans B-III et

$$T = \frac{1}{NL} \sum_{i \in S} \frac{y_i}{\alpha_i}$$

est hyper-admissible parmi toutes les stratégies P-sans biais de la moyenne par branche d'âge L.

Les 3 modèles des super-populations considérées recouvrent chacun un aspect des résultats de la partie A. Pour chacun des modèles, nous avons cherché la stratégie optimale parmi les stratégies sans biais de la moyenne $\bar{Y}_k$, $k = 1...L$

Nous nous sommes intéressés particulièrement au cas où il existe une bonne relation de régression entre la mesure sur un segment d'un âge donné et celles sur ses ramifications. On s'est inspiré alors des résultats du troisième modèle pour proposer une stratégie, mettant en valeur le plan

de sondage défini, pour l'estimation de la moyenne par arbre (ou par branche d'un certain âge fixé). Par simulation, nous comparons plusieurs estimateurs et nous constatons que le gain en efficacité peut atteindre 40 % quand on utilise une relation de régression.

Les programmes utilisés peuvent être repris dans d'autres problèmes, pour étudier la relation entre risque d'erreur et plan de sondage et choisir une stratégie. Cela suppose que l'on connait les valeurs approximatives des coefficients de régression et des variances résiduelles conditionnelles (déterminés par exemple à partir de l'année précédente). D'autres exemples d'utilisation peuvent être l'estimation de la biomasse ou de la concentration d'un produit chimique dans un peuplement d'arbres quelconque.

Ce travail peut-être poursuivi dans différentes directions, parmi lesquelles nous citons :

- lien entre les paramètres de dispersion $(\gamma_k)_{k=1...L}$ du plan de sondage et les variances résiduelles conditionnelles (dans la simulation, nous n'avons considéré qu'un plan de sondage "dispersé au maximum").
- considérer le cas du modèle obtenu à partir d'une combinaison des modèles utilisés (stratification, régression, loi de distribution connue).
- ne plus supposer que tout rameau d'âge k porte le même nombre de rameaux d'âge k-1 (hypothèse H4), ce qui nous permettrait alors d'étendre l'application de cette méthode à d'autres cas (voir la remarque R3 page 92).

D'autres suites pourraient être envisagées dans un but décisionnel, une fois déterminés des seuils de nuisibilité, en utilisant des techniques d'échantillonnage séquentiel.

A N N E X E A - II - 6

```
C TRANSFORMATION A APPLIQUER POUR AVOIR UN MODELE D' A.D.V. VERIFIANT :
C * L' INDEPENDANCE ENTRE MOYENNES ET VARIANCES
C * L' EGALITE DES VARIANCES
C A PARTIR DU TEST DE BARTLETT
C2FACTEURS EQUILIBRES
C       DIMENSION X(NI),Y(NI),K1(NI),K2(NI),YE(NS),EV(NS)
        DIMENSION X(378),Y(378),K1(378),K2(378),YE(63),EV(63)
        DIMENSION IND(17),D(78),NF(2),NIV(2),BT(20),TN(20)
        DOUBLE PRECISION EP(20),S2L
        CHARACTER*8 A(3)
C NUM VAR : NV
C NBRE IND : NI
C NBRE DE FACT : NFACT
C NUM FACTEURS : NF
C NBRE DE NIV DES FACT : NIV
        READ(5,59) NA
        DO 30 IK=1,NA
        REWIND(1)
        READ(5,60) A
        READ(5,59) NV,NI,NFACT,NX
        READ(5,59) (NF(I),I=1,NFACT),(NIV(I),I=1,NFACT)
        XN=FLOAT(NX)
        NS=NIV(1)
        IF(NFACT.EQ.2) NS=NS*NIV(2)
        NR=NI/NS
        C1=FLOAT(NS)/FLOAT(NR-1)-1./FLOAT(NS*(NR-1))
        WRITE(6,57) A
        WRITE(6,55) NS,C1,NR
        DO 21 I=1,4
        READ(1)
  21    CONTINUE
        DO 1 I=1,NI
        READ(1) IND,D
        X(I)=D(NV)
        K1(I)=IND(NF(1))
        IF(NFACT.EQ.2) K2(I)=IND(NF(2))
   1    CONTINUE
        DO 2 I=1,20
        ALPHA=FLOAT(10-I)/XN
        EP(I)=0.
        S2L=0.
        DO 22 J=1,NS
        EV(J)=0.
        YE(J)=0.
  22    CONTINUE
        DO 3 J=1,NI
        IF(ALPHA.EQ.0.) Y(J)=ALOG(X(J)+1.)
        IF(ALPHA.NE.0.) Y(J)=(X(J)+1.)**ALPHA
        L=K1(J)
        IF(NFACT.EQ.2) L=K1(J)+NIV(1)*(K2(J)-1)
        YE(L)=YE(L)+Y(J)
   3    CONTINUE
        DO 4 J=1,NS
        YE(J)=YE(J)/FLOAT(NR)
   4    CONTINUE
        DO 5 J=1,NI
        L=K1(J)
        IF(NFACT.EQ.2) L=K1(J)+NIV(1)*(K2(J)-1)
        EV(L)=EV(L)+(Y(J)-YE(L))**2
   5    CONTINUE
        DO 6 J=1,NS
        EV(J)=EV(J)/FLOAT(NR-1.
        EP(I)=EP(I)+EV(J)
        S2L=S2L+ALOG(EV(J))
   6    CONTINUE
        EP(I)=EP(I)/FLOAT(NS)
        BT(I)=FLOAT(NR-1)*(FLOAT(NS)*DLOG(EP(I))-S2L)
        TN(I)=SQRT(2.*BT(I))-SQRT(2.*FLOAT(NS)-3.)
   2    CONTINUE
        WRITE(6,50)
        DO 11 I=1,20
        ALP=(10-I)/XN
        IF(ALP.EQ.0.) WRITE(6,51) EP(I),BT(I),TN(I)
        IF(ALP.NE.0.) WRITE(6,52) ALP,EP(I),BT(I),TN(I)
  11    CONTINUE
        WRITE(6,53)
  30    CONTINUE
        STOP
  50    FORMAT(////T50,46(1H-)/T20,'   TRANSF',T50,1HI,2X,'S(VAR)  ',
     1 T65,1HI,'  BARTH',T80,1HI,'  T.NORM.',T95,1HI/T20,76(1H-))
  51    FORMAT(T20,1HI,'  X ----> LOG(X+1)',T50,1HI,3(F12.2,2X,1HI)/
     1 T20,1HI,T50,1HI,3(14X,1HI))
  52    FORMAT(T20,1HI,'  X ----> (X+1)**',F5.2,T50,1HI,3(F12.2,2X,1HI)/
     1 T20,1HI,T50,1HI,3(14X,1HI))
  53    FORMAT(T20,76(1H-))
  55    FORMAT(T20,' TEST DE BARTLETT : K=',I5,'    C1=',F5.2/
     + T40,'D.D.L : ',I5,' -1')
  57    FORMAT(1H1,T20,'VARIABLE :',A8,T50,'FACTEURS :',2A8)
  59    FORMAT(10I5)
  60    FORMAT(3A8)
        END
```

A N N E X E B - VI

```
      DIMENSION N(4,18),SC(5),BT(4,4),Y(4,1000),NS(4,300)
      DIMENSION THT(5),TB(5),TBA(5),TBB(5)
      DIMENSION BA(4,4),BB(4,4)
C
C     DIMENSIONS MAXIMALES :
C
       NDIM1=300
       NDIM2=18
C
C *   N = TABLEAU  DES TAILLES MAXIMALES POSSIBLES DES ECHANTILLONS
C         A COUT FIXE ; DONNES EN PARAMETRES ET DETERMINES PAR
C         LA SUBROUTINE TAILLE
C
C *   ESTIMATEURS :
C                 THT : HORVITZ - THOMPSON
C
C                 TB  : REGRESSION ; PARAMETRES BT DONNES
C
C                 TBA : REGRESSION ; PARAMETRES BT ESTIMES
C                                    MOINDRES CARRES (BT2)
C
C                 TBB : REGRESSION ; PARAMETRES BT ESTIMES
C                                    MOINDRES CARRES PONDERES (BT3)
C
C *   Y(I,J) = VALEUR DE LA VARIABLE SUR L'INDIVIDU J DE LA POPULATION I
C
C *   NS(I,J) = JEME INDIVIDU PRELEVE DE LA POPULATION I
C               DETERMINES PAR ECHANT
C
C *   PARAMETRES :
C     SC(1) : MOYENNE DE LA LOI DE LA POPULATION 1
C     SC(I+1),I=1,9 : VARIANCES DES RESIDUS DES POPULATIONS I
C
C     BT(I,J) : COEFFICIENT DE REGRESSION DE LA POPULATION J SUR
C               LA POPULATION I=1, I=2...J
C
C     L : AGE DE LA PLUS VIEILLE POPULATION
C     NL = EFFECTIF POPULATION L
C     M = NOMBRE DE RAMIFICATIONS D'AGE K PORTES PAR LE MEME K+1
C         POUR TOUT K=1,L-1
C     II = NOMBRE DE TAILLES DES ECHANTILLONS CORRESPONDANT
C         AU MEME COUT
C     NP = NOMBRE DE  POPULATIONS
C     NE = NOMBRE D'ECHANTILLONS
C
C     LECTURE DES PARAMETRES
      READ(5,100) L,NL,M,II,KR,NE,NP
       WRITE(6) L,NL,M,II,KR,NP,NE,NDIM1,NDIM2
       IF(II.GT.NDIM2) GO TO 1000
       NEST=NP*NE*II
       WRITE(7) NEST
      READ(5,101) (SC(I),I=1,L+2)
       WRITE(6) (SC(I),I=1,L+2)
      DO 90 J=2,L
       READ(5,101)(BT(I,J),I=1,J)
       WRITE(6) (BT(I,J),I=1,J)
   90 CONTINUE
       DO 61 K1=1,II
       READ(5,100) (N(I,K1),I=1,L)
       WRITE(6) (N(I,K1),I=1,L)
```

```
      IF(N(1,K1).GT.NDIM1) GO TO 1000
 61   CONTINUE
C
C     POUR CHAQUE TAILLE:
      DO 2 K1=1,II
C
C     GENERATION DE NE ECHANTILLON
      DO 2 J=1,NE
       CALL OLOAD('OV1  ',ISTAT)
      CALL ECHANT(L,NL,M,K1,N,NS)
C
C     GENERATION DE NP POPULATIONS POUR CHAQUE ECHANTILLON :
C
       DO 3 I=1,NP
       CALL OLOAD('OV2  ',ISTAT)
       CALL RAMIF(L,NL,M,KR,SC,BT,Y)
C
       CALL OLOAD('OV3  ',ISTAT)
       CALL MOYVAR(I,L,NL,M,Y,YM,VM)
       WRITE(6) YM,VM
C
C
       DO 16 I1=1,L+1
       THT(I1)=0.
       TBA(I1)=0.
       TBB(I1)=0.
 16    TB(I1)=0.
C
C 1- ESTIMATION DES PARAMETRES BT
C
       CALL OLOAD('OV4  ',ISTAT)
       CALL BT2(L,M,K1,KR,N,NS,Y,BA)
       CALL OLOAD('OV5  ',ISTAT)
       CALL BT3(L,M,K1,KR,N,NS,Y,BB)
C
C 2- CALCUL DES ESTIMATEURS
C
       CALL OLOAD('OV6  ',ISTAT)
       CALL PROJ3(L,M,K1,KR,N,NS,Y,BT,BA,BB,THT,TB,TBA,TBB)
C
       DO 66 I1=2,L
       WRITE(7) I,K1,J, (BA(J1,I1),BB(J1,I1),J1=1,I1)
 66    CONTINUE
C
       WRITE(6) (THT(I1),I1=1,L+1),(TB(I1),I1=1,L+1),(TBA(I1),I1=1,L+1),
     + (TBB(I1),I1=1,L+1)
C
  3    CONTINUE
  2    CONTINUE
  1   CONTINUE
C
 1000 STOP
C
  100 FORMAT(16I5)
  101 FORMAT(10F8.3)
      END
```

```
      SUBROUTINE RAMIF(L,NL,M,KR,SC,BT,Y)
C
C     GENERE NL ARBRES D' AGE L , TEL QUE CHAQUE RAMEAU D' AGE K
C     PORTE M RAMEAU D' AGE K-1; K=2,L
C
C     LA POP 1 SELON UNE LOI NORMALE DE MOYENNE SC(1)
C     LA POP K ,K=2,L , SELON UNE LOI NORMALE DE MOYENNE
C                       BT*(LA MOYENNE DES RAMEAUX PORTES)
C
C    LES VARIANCES RESIDUELLES CONDITIONNELLES DES LOIS
C    SONT EGALES A :
C                - SI KR=1 , SC(K+1)
C                - SI KR=2 , SC(K+1)*MOYENNE**2
C
      DIMENSION SC(5),BT(4,4),Y(4,1000),G(5,5)
C
      IP=12
      S1=SC(1)
      DO 1 I1=1,L
      S3=SQRT(SC(I1+1)*12./FLOAT(IP))
      N1=M**(L-I1)
      DO 1 I2=1,NL*N1
      IF(I1.EQ.1) GO TO 2
      S1=BT(1,I1)
      DO 3 I3=2,I1
      YM=0.
      NJ=M**(I1-I3+1)
      DO 4 I4=(I2-1)*NJ+1,I2*NJ
 4    YM=YM+Y(I3-1,I4)
 3    S1=S1+BT(I3,I1)*YM/FLOAT(NJ)
 2    Y(I1,I2)=0.
      DO 5 I5=1,IP
      Y(I1,I2)=Y(I1,I2)+RANF(-1)
 5    CONTINUE
C CAS E.T = CTE
      Y(I1,I2)=(Y(I1,I2)-FLOAT(IP)/2.)*S3+S1
C CAS E.T = CTE * MOY
      IF(KR.EQ.2) Y(I1,I2)=(Y(I1,I2)-S1)*ABS(S1)+S1
 1    CONTINUE
 10   CONTINUE
      RETURN
      END
```

```
      SUBROUTINE ECHANT(L,NL,M,K1,N,NS)
C
C PRELEVE N(K,K1) RAMEAUX DE LA POP K , K=1,L
C DE LA FACON LA PLUS DISPERESEE POSSIBLE
C ET EN TENANT COMPTE DE L' INCLUSION DES ECHANTILLONS
C
      DIMENSION N(4,18),NS(4,300)
C
C POPULATION L
C
    9 IJ=0
      P=FLOAT(N(L,K1))/FLOAT(NL)
      DO 1 I1=1,NL
      X=RANF(-1)
      IF(X.GT.P) GO TO 1
      IJ=IJ+1
      NS(L,IJ)=I1
    1 CONTINUE
      IF(IJ.NE.N(L,K1)) GO TO 9
      IF(L.EQ.1) GO TO 99
C
C   POP I2 , I2=L-1....1
C
      DO 50 I2=L-1,1,-1
C    SI UN INDIVIDU J DE I2+1 EST PRELEVE , ALORS TOUS LES INDIVIDUS
C    DE I2 EN RELATION AVEC J LE SONT AUSSI
      DO 51 I3=1,N(I2+1,K1)
      DO 51 I4=1,M
      I5=(NS(I2+1,I3)-1)*M+I4
      I6=(I3-1)*M+I4
      NS(I2,I6)=I5
   51 CONTINUE
C
C    POUR LES AUTRES INDIVIDUS DE I2 :
C -1
C   CALCUL DES PROBABILITES D' INCLUSION
C
        J1=N(I2,K1)-M*N(I2+1,K1)
        K2=NL*M**(L-I2)-M*N(I2+1,K1)
        P=FLOAT(J1)/FLOAT(K2)
      IF(P.LT.1.) GO TO 54
      NJ=M**(L-I2)
      DO 52 I3=1,NL*NJ
      NS(I2,I3)=I3
   52 CONTINUE
      GO TO 50
   54 Q1=1.
      DO 53 K=1,L-I2
      Q1=Q1/FLOAT(M)
      IF(P.LT.Q1) GO TO 53
      DO 70 J1=M-1,1,-1
      Q3=J1*Q1
      IF(P.LT.Q3) GO TO 70
      JJ=J1
      Q2=P/Q1-Q3
      IK=K+I2
      Q4=FLOAT(JJ)/FLOAT(M)
      Q5=M*Q1
      GO TO 56
   70 CONTINUE
```

```
 53     CONTINUE
        IK=L
        JJ=0
        Q3=0.
 56     JK=M*N(I2+1,K1)
        N1=M**(L-IK)
        N2=M**(IK-I2-1)
        NK=M**(IK-I2)
        JB=0
        DO 57 J4=1,NL*N1
        DO 58 J5=1,NK
        J6=(J4-1)*NK+J5
        DO 58 J7=1,JK
        IF(NS(I2,J7).EQ.J6) GO TO 57
 58     CONTINUE
        JB=JB+1
 57     CONTINUE
        Q2=FLOAT(N(I2,K1)-N(I2+1,K1)*M)/FLOAT(JB)-FLOAT(JJ)
 62     IF(Q2.LT.1.) GO TO 55
        JJ=JJ+1
        IF(JJ.EQ.M) GO TO 61
        Q3=JJ*Q1
        Q4=FLOAT(JJ)/FLOAT(M)
        Q2=Q2-1.
        GO TO 62
 61     IK=IK+1
        IF(IK.GT.L) GO TO 63
        JJ=1
        Q5=Q1
        Q1=Q1/FLOAT(M)
        Q3=Q1
        Q4=1./FLOAT(M)
        GO TO 56
 63     JK=M*N(I2+1,K1)
        N1=M**(L-I2)
C
C 2A-   SI LES TAILLES DES ECHANTILLONS SONT GRANDES
C
        DO 64 I3=1,NL*N1
C    LES RAMEAUX I3 DE I2 N' ONT PAS ETE DEJA PRELVES
        DO 65 J4=1,JK
        IF(NS(I2,J4).EQ.I3) GO TO 64
 65     CONTINUE
        X=RANF(-1)
        IF(X.GT.P) GO TO 64
        JK=JK+1
        NS(I2,JK)=I3
 64     CONTINUE
        IF(JK.NE.N(I2,K1)) GO TO 63
        GO TO 99
C
C 28-   SINON :
C
 55     JK=M*N(I2+1,K1)
C  POUR TOUT RAMEAU J4 DE IK :
        DO 71 J4=1,NL*N1
C  28-1- J4 NE PORTE AUCUN RAMEAU I2 PRELEVE
        DO 72 J5=1,NK
        J6=(J4-1)*NK+J5
        DO 72 J7=1,JK
```

```
      IF(NS(I2,J7).EQ.J6) GO TO 71
 72   CONTINUE
C 28-2- PRELEVER JJ RAMEAUX SUR J4
      IF(JJ.EQ.0) GO TO 73
      IF(JJ.EQ.1.OR.(IK+1).EQ.I2) GO TO 74
C 28-2-A- CAS JJ>1 ET IK+1 > I2
C     PRELEVER JJ RAMEAUX IK+1 SUR CHAQUE J4, PUIS
C     SUR CHAQUE IK+1 PRELEVE, TIRER UN RAMEAU I2
 75   J5=0
      J9=JK
      DO 76 I3=1,M
      I4=(J4-1)*M+I3
      X=RANF(-1)
      IF(X.GT.Q4) GO TO 76
      J5=J5+1
 77   J6=0
      DO 78 I5=1,N2
      I6=(I4-1)*N2+I5
      DO 79 I7=1,J9
      IF(NS(I2,I7).EQ.I6) GO TO 78
 79   CONTINUE
      X=RANF(-1)
      IF(X.GT.Q5) GO TO 78
      J6=J6+1
      NS(I2,J9+J6)=I6
 78   CONTINUE
      IF(J6.NE.1) GO TO 77
      J9=J9+1
 76   CONTINUE
      IF(J5.NE.JJ) GO TO 75
      JK=JK+JJ
      GO TO 73
C 28-2-B- CAS JJ=1 OU IK=I2+1
 74   J5=0
      J9=JK
      DO 81 I5=1,NK
      I6=(J4-1)*NK+I5
      DO 82 I7=1,J9
      IF(NS(I2,I7).EQ.I6) GO TO 81
 82   CONTINUE
      X=RANF(-1)
      IF(X.GT.Q3) GO TO 81
      J5=J5+1
      NS(I2,J9+J5)=I6
 81   CONTINUE
      IF(J5.NE.JJ) GO TO 74
      JK=JK+JJ
C 28-3- SUR CHAQUE J4 ,PRELEVER AU PLUS UN I2
 73   IF(Q2.EQ.0.) GO TO 71
      X=RANF(-1)
      IF(X.GT.Q2) GO TO 71
 83   J5=0
      J9=JK
      DO 84 I5=1,NK
      I6=(J4-1)*NK+I5
      DO 85 I7=1,J9
      IF(NS(I2,I7).EQ.I6) GO TO 84
 85   CONTINUE
      X=RANF(-1)
      IF(X.GT.Q1) GO TO 84
```

```
      J5=J5+1
      NS(I2,J9+J5)=I6
84    CONTINUE
      IF(J5.NE.1) GO TO 83
      JK=JK+J5
71    CONTINUE
      IF(JK.NE.N(I2,K1)) GO TO 55
50    CONTINUE
99    CONTINUE
      RETURN
      END
```

```
      SUBROUTINE MOYVAR(I,L,NL,M,Y,YM,YV)
C
C CALCULE MOYENNE ET VARIANCE DE LA POP SUR ARBRES
C
      DIMENSION Y(4,1000),T(4)
      YM=0.
      YV=0.
C
      DO  1 I1=1,NL
      TX=Y(L,I1)
      DO  2 I2=1,L-1
      NM=M**(L-I2)
      DO  2 I3=1,NM
      I4=(I1-1)*NM+I3
      TX=TX+Y(I2,I4)
2     CONTINUE
      YM=YM+TX
      YV=YV+TX**2
1     CONTINUE
      TX=YM**2/FLOAT(NL)
      YM=YM/FLOAT(NL)
      YV=(YV-TX)/FLOAT(NL-1)
      RETURN
      END
```

```
      SUBROUTINE BT2(L,M,K1,KR,N,NS,Y,BT)
      DIMENSION N(4,18),BT(4,4),Y(4,1000),NS(4,300),X(300,4)
      DIMENSION XM(4),B(4)
      DOUBLE PRECISION X
C
C     ESTIMATION DES PARAMETRES A PARTIR DE L' ECHANTILLON
C     SUR VARIABLES CENTREES
C
      DO 15 I5=2,L
        NR=N(I5,K1)
        NC=I5
        DO 61 I1=1,I5
 61     XM(I1)=0.
        DO 52 I6=1,NR
          J9=NS(I5,I6)
          X(I6,NC)=Y(I5,J9)
          XM(NC)=XM(NC)+X(I6,NC)
          DO 52 I7=1,I5-1
            X(I6,I7)=0.
            MM=M**(I5-I7)
          DO 54 J1=1,MM
             I8=(J9-1)*MM+J1
              X(I6,I7)=X(I6,I7)+Y(I7,I8)
 54   CONTINUE
            X(I6,I7)=X(I6,I7)/FLOAT(MM)
          XM(I7)=XM(I7)+X(I6,I7)
 52       CONTINUE
      DO 60 I1=1,I5
      XM(I1)=XM(I1)/FLOAT(NR)
      DO 62 I2=1,NR
      X(I2,I1)=X(I2,I1)-XM(I1)
 62   CONTINUE
 60   CONTINUE
C
          CALL REGRES(X,B,NR,NC)
C
 51       BT(1,I5)=XM(I5)
          DO 53 I6=2,I5
          BT(1,I5)=BT(1,I5)-XM(I6-1)*B(I6-1)
 53       BT(I6,I5)=B(I6-1)
C
 15   CONTINUE
 14   CONTINUE
      RETURN
      END
```

```
      SUBROUTINE BT3(L,M,K1,KR,N,NS,Y,BT)
      DIMENSION N(4,18),BT(4,4),Y(4,1000),NS(4,300),B(5),X(300,5)
      DOUBLE PRECISION X
C
C     ESTIMATION DES PARAMETRES A PARTIR DE L' ECHANTILLON
C
C       MOINDRES CARRES PONDERES
C
      DO 15 I5=2,L
        NR=N(I5,K1)
        NC1=I5
        NC=I5+1
        DO 52 I6=1,NR
          X(I6,1)=1.
          J9=NS(I5,I6)
          X(I6,NC)=Y(I5,J9)
          DO 52 I7=2,I5
            J7=I7-1
            X(I6,I7)=0.
            M1=M**(I5-J7)
              DO 54 J1=1,M1
              I8=(J9-1)*M1+J1
 54           X(I6,I7)=X(I6,I7)+Y(J7,I8)
            X(I6,I7)=X(I6,I7)/FLOAT(M1)
 52       CONTINUE
C
      DO 1 I=1,NR
      DO 2 J=1,NC
      IF(J.EQ.NC1) GO TO 2
      IF(X(I,NC1).NE.0) X(I,J)=X(I,J)/X(I,NC1)
 2    CONTINUE
      X(I,NC1)=1.
 1    CONTINUE
          CALL REGRES(X,B,NR,NC)
C
          DO 53 I6=1,I5
          BT(I6,I5)=B(I6)
 53       B(I6)=0
 15   CONTINUE
      RETURN
      END
```

```
      SUBROUTINE PROJ3(L,M,K1,KR,N,NS,Y,BT,BA,BB,THT,TB,TBA,TBB)
C
C     CALCUL DES ESTIMATEURS
C
      DIMENSION N(4,18),NS(4,300),Y(4,1000),THT(5),TB(5),TBA(5),TBB(5)
      DIMENSION BT(4,4),BA(4,4),BB(4,4)
      DIMENSION IU(300),IV(300),T(3)
C
      DO 16 I1=1,L+1
      THT(I1)=0.
      TB(I1) =0.
      TBA(I1)=0.
   16 TBB(I1)=0.
C
C   HORVITZ - THOMPSON
C
      DO 12 I1=2,L+1
      M1=M**(L-I1+1)
      I2=I1-1
      J2=N(I2,K1)
      DO 13 J1=1,J2
      J3=NS(I2,J1)
   13 THT(I1)=THT(I1)+Y(I2,J3)
      THT(I1)=THT(I1)/FLOAT(J2)
   12 THT(1)=THT(1)+THT(I1)*M1
C
      TB(2)=THT(2)
      TBA(2)=THT(2)
      TBB(2)=THT(2)
      M1=M**(L-1)
      TB(1)=TB(2)*M1
      TBA(1)=TB(1)
      TBB(1)=TB(1)
C
C   AUTRES ESTIMATEURS
C
      IF(L.EQ.1) GO TO 1
      DO 1 I1=2,L
      NV=0
      NU=N(I1,K1)
      DO 2 J1=1,NU
    2 IU(J1)=NS(I1,J1)
      CALL SSECH(1,I1,M,K1,N,NS,IV,NV)
      CALL PROJ4(I1,M,K1,KR,N,NS,Y,IU,NU,IV,NV,BT,BA,BB,T)
C
      M1=M**(L-I1)
      TB(I1+1)=T(1)
      TB(1)=TB(1)+T(1)*M1
      TBA(I1+1)=T(2)
      TBA(1)=TBA(1)+T(2)*M1
      TBB(I1+1)=T(3)
      TBB(1)=TBB(1)+T(3)*M1
    1 CONTINUE
      RETURN
      END
```

```
      SUBROUTINE PROJ4(L,M,K1,KR,N,NS,Y,IU,NU,IV,NV,BT,BA,BB,T)
C
C   ESTIMATION DU TOTAL SUR LES BRANCHES D' AGE  L , NON PRELEVEES,
C   MAIS PORTANT CHACUNE AU MOINS UN RAMEAU DE 1 AN QUI A ETE PRELEVE
C   DE MANIERE RECURSIVE, EN UTILISANT LES COEFF DE REGRESSION BT,BA,BB
C
      DIMENSION N(4,18),NS(4,300),Y(4,1000),IU(300),IV(300)
      DIMENSION BT(4,4),BA(4,4),BB(4,4),T(3)
      DIMENSION PT(4,3),JL(4),JNU(4),JNV(4),JK(4),JJ(4)
      DIMENSION JU(4,300),JV(4,300),IK(300),IL(300)
      IH=0
    1 TL=0.
      DO 11 I1=1,NU
      J1=IU(I1)
   11 TL=TL+Y(L,J1)
      DO 21 I1=1,3
   21 T(I1)=TL
      IF(L.EQ.1.OR.NV.EQ.0) GO TO 10
      T(1)=T(1)+NV*BT(1,L)
      T(2)=T(2)+NV*BA(1,L)
      T(3)=T(3)+NV*BB(1,L)
      K=1
    3 J=1
    2 JX=IV(J)
      CALL PROJ1(1,K,L,JX,M,K1,N,NS,IK,NK,IL,NL)
C PRESERVATION DES PARAMETRES
      IH=IH+1
      DO 22 I1=1,3
   22 PT(IH,I1)=T(I1)
      JL(IH)=L
      JNU(IH)=NU
      JNV(IH)=NV
      DO 12 I1=1,NU
   12 JU(IH,I1)=IU(I1)
      DO 13 I1=1,NV
   13 JV(IH,I1)=IV(I1)
      JK(IH)=K
      JJ(IH)=J
C APPEL
      L=K
      NU=NK
      NV=NL
      DO 14 I1=1,NU
   14 IU(I1)=IK(I1)
      DO 15 I1=1,NV
   15 IV(I1)=IL(I1)
      GO TO 1
   10 DO 24 I1=1,3
      IF((NU+NV).EQ.0) T(I1)=0.
      IF((NU+NV).GT.0) T(I1)=T(I1)/FLOAT(NU+NV)
   24 CONTINUE
      IF (IH.EQ.0) RETURN
C RESTITUTION
      L=JL(IH)
      NU=JNU(IH)
      NV=JNV(IH)
      DO 16 I1=1,NU
   16 IU(I1)=JU(IH,I1)
      DO 17 I1=1,NV
   17 IV(I1)=JV(IH,I1)
```

```
      K=JK(IH)
      J=JJ(IH)
      T(1)=PT(IH,1)+BT(K+1,L)*T(1)
      T(2)=PT(IH,2)+BA(K+1,L)*T(2)
      T(3)=PT(IH,3)+BB(K+1,L)*T(3)
      IH=IH-1
      J=J+1
      IF(J.LE.NV) GO TO 2
      K=K+1
      IF(K.LE.L-1) GO TO 3
      GO TO 10
      END
```

```
      SUBROUTINE REGRES(X,E,NR,NC1)
C
C   CALCUL DES COEFFICIENTS DE REGRESSION PAR LA METHODE
C   DE GRAM-SCHMIDT MODIFIEE
C
      DIMENSION X(300,4),Z(4,5),E(4),D(4)
      DOUBLE PRECISION D,SXY,X
      ATOM=0.00001
      NC=NC1-1
      DO 2 I=1,NC
      DO 1 J=1,NC1
    1 Z(I,J)=0
    2 Z(I,I)=1
      DO 10 J=1,NC
      D(J)=0
      DO 6 I=1,NR
    6 D(J)=D(J)+X(I,J)*X(I,J)
      IF(D(J).LT.ATOM) GO TO 12
      DO 10 L=J+1,NC1
      SXY=0
      DO 7 I=1,NR
    7 SXY=SXY+X(I,J)*X(I,L)
      B=SXY/D(J)
      DO 8 I=1,NR
    8 X(I,L)=X(I,L)-B*X(I,J)
      DO 9 K=1,NC
    9 Z(K,L)=Z(K,L)-B*Z(K,J)
   10 CONTINUE
      DO 11 J=1,NC
   11 E(J)=-Z(J,NC1)
      RETURN
   12 NC1=-J
      RETURN
      END
```

```
      FUNCTION RANF (SEED)
C
C     M(P)=TABLE OF P PREVIOUS RANDOM NUMBERS.
C     P,Q,=POLYNOMIAL PARAMETERS:X**P+X**Q+1
C     .NOT. OPERATOR IMPLEMENTED IN ARITHMETIC.
C     INTSIZ=INTEGER SIZE (BITS) OF HOST MACHINE: E.G.,
C     UNIVAC 35BITS, PDP 11 15BITS.
C
      LOGICAL AA,BB,AATEMP,FIRST
      INTEGER A,B,P,Q,INTSIZ,M(98),SEED
      EQUIVALENCE (AA,A),(BB,B)
      DATA P,Q,INTSIZ/98,27,15/
      DATA J/0/
      DATA AM/32767./
      DATA FIRST/.TRUE./
      DATA M/5283,9027,29649,17902,27112,
     &17080,2628,15041,23716,22936,
     &9924,24694,9764,22098,15973,
     &10058,15791,4269,4988,5600,
     &9895,29140,9438,12901,4459,
     &4568,32651,15347,5953,17715,
     &31041,5097,8630,1904,32107,
     &30666,30512,1269,24281,9867,
     &29531,12130,10381,27034,20216,
     &28531,5443,16203,9383,3995,
     &10742,12168,2719,25050,23493,
     &20369,14351,29871,28663,31034,
     &26109,32629,9883,23742,18464,
     &27361,27768,11472,5273,8780,
     &25038,21044,24686,8285,19925,
     &22532,22381,28110,25430,3311,
     &10455,13764, 494,12050,10723,
     &16496,23988,4905,9832,13183,
     &20579,1287,29843,14365,15025,
     &27377,26009,19133/
      IF(.NOT.FIRST)GO TO 20
      IF(SEED.LE.1)GO TO 15
      FIRST=.FALSE.
      DO 10 I=1,SEED
      J=J+1
      IF(J.GT.P)J=1
      K=J+Q
      IF(K.GT.P)K=K-P
      A=M(J)
      B=M(K)
      AATEMP=AA.OR.BB
      BB=AA.AND.BB
      AA=AATEMP
      M(J)=A-B
      RANF=FLOAT(M(J))/AM
   10 CONTINUE
      RETURN
   15 FIRST=.FALSE.
   20 J=J+1
      IF(J.GT.P)J=1
      K=J+Q
      IF(K.GT.P)K=K-P
      A=M(J)
      B=M(K)
      AATEMP=AA.OR.BB
```

```
      BB=AA.AND.BB
      AA=AATEMP
      M(J)=A-B
      RANF=FLOAT(M(J))/AM
      RETURN
      END
```

```
      SUBROUTINE PROJ1(J,K,L,I,M,K1,N,NS,JU,MU,JV,MV)
C
C  JU EST L' ENSEMBLE DES MU RAMEAUX D' AGE K PORTES PAR LA BRANCHE I
C        D' AGE L
C
C  JV EST L ' ENSEMBLE DES MV RAMEAUX D ' AGE K , PORTES PAR LA BRANCHE I
C        D' AGE L , NON PRELEVES , ET PORTEURS CHACUN D' AU MOINS
C        UN RAMEAU D ' AGE J QUI A ETE PRELEVE
C
      DIMENSION N(4,18),NS(4,300),JU(300),JV(300),IW(300)
      DIMENSION IU(300),IV(300)
      MU=0
      MV=0
      IF(K.GT.L) GO TO 1
      NW=M**(L-K)
C   IW : RAMEAUX D' AGE K PORTES PAR LA BRANCHE I D' AGE L
      DO 2 I1=1,NW
      IW(I1)=(I-1)*NW+I1
  2   CONTINUE
      NU=N(K,K1)
C   IU : RAMEAUX D' AGE K PRELEVES
      DO 3 I1=1,NU
      IU(I1)=NS(K,I1)
  3   CONTINUE
      CALL PROJ2(IW,NW,IU,NU,JU,MU)
      IF(J.GE.K) GO TO 1
      CALL SSECH(J,K,M,K1,N,NS,IV,NV)
      CALL PROJ2(IW,NW,IV,NV,JV,MV)
  1   CONTINUE
      RETURN
      END
```

```
C
C   TAILLES DES ECHANTILLONS A PRELEVER DES DIFFERENTES POPULATIONS
C   ETANT DONNES LES PARAMETRES DE LA FONCTION COUT : C ET G
C
C   TAILLES MAXIMALES POUR CHAQUE POPULATION
C
C
      DIMENSION N(4,4),NN(4)
      DIMENSION IL(4),PC(4),IK(4),IJ(4),II(4)
C
       READ(5,81) LL,NL,M,C,G
       WRITE(6,80) LL,NL,M,C,G
C
      IF(LL.LT.1.OR.LL.GT.4) GO TO 39
       ML=M*LL-LL+M
       GM=G*M
       GG=G/(G-1.)
       GX=GM**(LL-1)
       NM=INT(C/(GX*M))+1
       WRITE(6,85) NM
       IF(NM.GE.NL) GO TO 21
      DO 25 J=1,LL
      DO 25 K=1,LL
      N(J,K)=0
  25  CONTINUE
      IF(LL.GT.1) GO TO 31
      I1=INT((1.-1./G)*C)
      IF(I1.LT.1) GO TO 20
      K1=1
      N(1,1)=I1
      GO TO 22
  31   L=LL
       IH=0
       MJ=M**(L-1)
       K1=INT(FLOAT(MJ*ML-1)/FLOAT(M-1))
       J1=NL*MJ
   3   CONTINUE
       I=K1
   6   CONTINUE
       NN(LL-L+1)=I
       IF(L.EQ.LL) CC=C-GG*I
       IF(L.LT.LL) CC=C-GM**(LL-L)*I
       IF(L.NE.1) GO TO 4
       IF(ABS(CC).GT.1) GO TO 7
       WRITE(6,*) (NN(J),J=1,LL)
       DO 5 J=1,LL
       IF(NN(J).LT.N(J,J)) GO TO 5
       DO 5 K=1,LL
       N(K,J)=NN(K)
   5   CONTINUE
       GO TO 7
   4  IH=IH+1
       IL(IH)=L
       PC(IH)=C
       IK(IH)=K1
       IJ(IH)=J1
       II(IH)=I
       L=L-1
       C=CC
       K1=INT(FLOAT(M**(L-1)*ML-1)/FLOAT(M-1))
```

```
      J1=INT(FLOAT(I-1)/FLOAT(M))
      GO TO 3
 7    I=I+1
      IF(I.LE.J1) GO TO 6
100   CONTINUE
      IF(IH.EQ.0) GO TO 10
      L=IL(IH)
      C=PC(IH)
      K1=IK(IH)
      J1=IJ(IH)
      I=II(IH)
      IH=IH-1
      GO TO 7
 10   CONTINUE
      KL=0
      DO 23 J=1,LL
      IF(N(J,1).EQ.0) GO TO 20
23    CONTINUE
      KL=LL
      I=2
 30   DO 24 KM=1,I-1
      DO 27 J=1,LL
      IF(N(J,I).NE.N(J,KM)) GO TO 24
27    CONTINUE
      GO TO 35
 24   CONTINUE
      GO TO 28
 35   KL=KL-1
      IF(I.GT.KL)GO TO 28
      DO 29 J=I,KL
      DO 29 K=1,LL
      N(K,J)=N(K,J+1)
29    CONTINUE
      GO TO 30
28    I=I+1
      IF(I.LE.KL) GO TO 30
      WRITE(6,83)
      DO 82 I=1,KL
      WRITE(6,*) (N(J,I),J=1,LL)
82    CONTINUE
      GO TO 22
20    WRITE(6,*) ' COUT NON VALABLE'
     GO TO 22
 21   WRITE(6,*) '  COUT TROP GRAND '
      GO TO 22
39    WRITE(6,*) ' AGE NON VALABLE'
 22   CONTINUE
      STOP
80    FORMAT(T10,'    AGE :',I5/T10,' NOMBRE :',I5/T10,'  RAMIF :',
     + I5/T10,'   COUT :',F8.1/T10,'  COEFF :',F8.1////
     + '  TOUTES LES TAILLES POSSIBLES :'/)
81    FORMAT(3I5,2F8.1)
83    FORMAT(//'   TAILLES EXTREMES POSSIBLES :'//)
 85   FORMAT( /T10,'TAILLE MAXIMALE BRANCHES D AGE L : ',I5)
     END
```

```
      SUBROUTINE PROJ2(IU,NU,IV,NV,ID,ND)
C
C     INTERSECTION DE 2 ENSEMBLES
C
      DIMENSION IU(300),IV(300),ID(300)
      ND=0
      IF(NU.EQ.0.OR.NV.EQ.0) GO TO 1
      DO 1 I1=1,NU
      DO 2 I2=1,NV
      IF(IU(I1).NE.IV(I2)) GO TO 2
      ND=ND+1
      ID(ND)=IV(I2)
    2 CONTINUE
    1 CONTINUE
      RETURN
      END
```

```
      SUBROUTINE SSECH(K,L,M,K1,N,NS,IV,NV)
      DIMENSION N(4,18),NS(4,300),IV(300)
      NV=0
      IF(K.EQ.L) GO TO 1
      X2=FLOAT(M**(L-K))
      DO 1 I=1,N(K,K1)
      X1=FLOAT(NS(K,I)-1)
      J1=INT(X1/X2)+1
      DO 2 J=1,N(L,K1)
      IF(J1.EQ.NS(L,J)) GO TO 1
    2 CONTINUE
      IF(NV.EQ.0) GO TO 3
      DO 4 I9=1,NV
      IF(J1.EQ.IV(I9)) GO TO 1
    4 CONTINUE
    3 NV=NV+1
      IV(NV)=J1
    1 CONTINUE
      RETURN
      END
```